1

Tales and Woes of High Frequency Trading: an Introduction

René A. Carmona[1]

Bendheim Center for Finance
Department of Operations Research & Financial Engineering,
Princeton University, Princeton, NJ 08544, USA
email: rcarmona@princeton.edu

1.1 Introduction

1.1.1 Standard Assumptions in Finance

One of the basic assumptions of the early mathematical theory of financial markets is the absence of friction and consequently, the fact that securities have one price (*law of one price*), and that they can be sold and bought at this price in any desired quantity. This form of infinite liquidity implies that the sizes and the frequencies of the transactions have no impact on the prices at which the transactions take place. This lack of price impact is often justified by limiting the scope of these models to the behavior of so-called *small investors*. While this restriction can exonerate the models of the lack of price impact, it cannot justify the side effect of infinite liquidity.

As demonstrated by the recent financial crisis, the existence of a quoted price is not enough for transactions to be possible, and to actually occur. Financial agents willing to sell and buy will have to agree on a price level, a quantity, and a specific timing for the transaction to take place.

The above claims should be understood as a zealous invitation to the study of market microstructure. This invitation does not imply that frictions have not been studied mathematically. Indeed, transaction costs (at least proportional transaction costs) have not been included in mathematical models for almost twenty years, and many attempts have been made to extend Merton's theory of optimal portfolio choice in order to capture their impact. We refer the reader to [39, 44, 91] for the earliest contributions, and to Muhle-Karbe's lectures included in this volume and expanded in [71], for recent developments and an up-to-date set of references.

[⋆] These notes cover the first five lectures of a ten lecture course on high frequency markets at the first Princeton RTG Summer school on Financial Mathematics from June 21 to June 30, 2013. The last five lectures were given by A. Schied. They were based on advanced material already published in [90], [9] and [8], so closely that A. Schied thought it was neither appropriate nor necessary to provide still another exposé of the material.
This research was partially supported by NSF DMS-0456195.

Similarly, attempts have been made to include liquidity frictions in mathematical models. For example, [35] assimilates liquidity frictions to a form of added costs to a transaction. While remaining an intuitively appealing way to include illiquidity issues in the models, this approach is still unsatisfactory for large trades (especially over short periods of time) and High Frequency Trading (HFT).

Our mild introduction to high frequency trading is aimed at mathematicians and financial engineers curious about the most common forms of electronic trading. Among many other things, we shall not consider the thread of literature on ”agent based models”. See for example [94]. Still, this introduction is screaming for the analysis of the market microstructure in order to understand how and why are buy and sell orders are posted and executed.

1.1.2 Traditional Markets

In order to understand trading on the electronic markets and especially limit order book trading, it is important to review the typical characteristics of the traditional trading processes before the advent of the electronic markets. Most financial trades used to take place in *quote - driven markets* where individual buy and sell orders are collected by *dealers* and/or *market makers* who in turn publish the prices at which they are subsequently willing to buy and sell the financial interest being traded. This is the case for example in the New York Stock Exchange (NYSE) specialist system. In order for the business of market making to make sense, the sell price published needs to be higher than the buy price. The difference (often called the *spread*) is the source of profit compensating the market maker for taking risk in making the market. Market makers are considered as *liquidity providers* and in these lectures, we shall use interchangeably the terms of *market maker* and *liquidity provider*. Typically, the other market participants have only access to the prices published by the market-makers, and their decisions to buy or sell are based on these prices.

For the sake of the following discussion we specify the meaning given to an expression which will be used often in the sequel: *adverse selection* is a term frequently used in the economic literature to describe the quandary of a market maker facing agents who have better information about the value of the financial interest being traded, and who can therefore make a profit by buying or selling, often repeatedly, with the market-maker [83]).

Liquidity is also a word often used despite the fact that it is rather difficult to define. In these notes, we shall often use this word to qualify the level of friction in a market, but also to identify certain types of agents by identifying forms of trading behaviors. Case in point, some of the market models we discuss below are based on the interactions between two families of traders whom we distinguish as liquidity providers and liquidity takers.

1.1.3 New Markets

A second category of trading venue can be characterized as *order-driven markets*. These are typically electronic platforms which aggregate all the available orders in a Limit Order Book (LOB). These markets include the New York Sock Exchange (NYSE), NASDAQ and the London Stock Exchange (LSE). One of the special features of these markets is that the same stock can be traded on several venues. This makes price discovery more difficult, especially because most instruments can also be traded off market, without printing the trades to a publicly accessible data source like the so-called consolidated tape. As advocated by the proponents of the existence of, and the competition between multiple exchanges, the resulting alternatives lead to lower fees and smaller tick sizes. However, agents will have to decide how to split and route their orders to best use the competing exchanges, relying more and more on on hardware and software called *smart order routers* or SORs. See for example [42] for an attempt at addressing the problem of optimal liquidation of a position when several venues can be accessed by the (electronic) broker.

Even though we will not spend much time discussing these alternatives, we shall also mention below the so-called dark pools. See nevertheless [75] for complements.
.

1.1.4 High Frequency Trading

High-frequency trading (HFT) is a specific form of algorithmic trading. Other forms of algorithmic trading are implemented at low frequency. We discuss some of these applications in Section 1.5. It implements proprietary trading strategies to move in and out of positions in fractions of a second. Instead of relying on macro-economic factors, these strategies depend on the detection of unusual patterns of market activities and price anomalies, and their success is mostly based on the speed of execution, hence the reliance on sophisticated computer systems, and state of the art communication systems to facilitate access to the electronic markets.

It is often claimed that HFT accounts for 60 – 75% of all share volume on many exchanges. While speculative, these figures are commonly accepted as they sound plausible. Moreover, 10% of that is qualified as *"predatory"*, which represents approximately 600 million shares per day. At \$0.01-\$0.02 per share, predatory HFT is profiting \$6-\$12 million a day or \$1.5-\$3 billion per year.

Predatory trading, while possibly undesirable or unethical, is perfectly legal. It has to be differentiated from front-running which is the illegal practice of using insider information to trade ahead of customer orders. In the context of HFT, however, this term is also used to describe firms using their speed advantage to trade before slower participants.

Algorithmic trading is a source of serious concern. The most obvious are based on the fact that trading firms are moving their computing facilities *closer* to the trading platform to avoid become victims of *latency* arbitrage, and increasingly rely on benchmark tracking execution algorithms. Latency is the delay between the transmission of information from a source and the reception of the signals at destination, typically the time between the placement of an order on an electronic trading system, and the actual execution of that order. This is the source of a new phenomenon called *co-location*. This practice was made possible by exchanges or trading venues renting out space in close physical proximity to their trading servers. Trading firms use this option to locate their computer systems near or at the exchanges systems to reduce the distance traveled by their trading signals.

1.1.5 Pros & Cons of High Frequency Trading

The arguments most often used in favor of high frequency trading are based on the fact that 1) high speed trading has reduced the size of the tick separating quotes, and hence, the cost of transactions; 2) high frequency traders provide extra liquidity to the market; 3) Dark pools reduce trade execution costs from price impact; 4) the markets are more efficient.

However, detractors of HFT claim that 1) it is the source of a needless and expensive technological *arms race*; 2) dark trading incentivizes price manipulation, fishing and predatory trading; 3) the fact that little or no human oversight is possible increases systemic risk (see for example the reports [36] or [63] for a forensic analysis of what is believed to be behind the flash crash of May 6, 2010); 4) high frequency trading algorithms do not use *economic fundaments* (e.g. value & profitability of a firm).

1.1.6 Some Highly Publicized Mishaps

The most significant mishap of high frequency trading is without the shadow of a doubt the ndex[sub]flash crash *Flash Crash of May 6, 2010*. On that day, the Dow Jones IA index plunged almost 1000 points, though ti recovered in minutes, the biggest one-day point decline (998.5 points to be specific). According to the many ensuing investigations, at 2:32 pm on that day, a *mutual fund program* started to sell $75,000$ E-Mini S&P 500 contracts (≈ 4.1 billion USD) at an execution rate of 9%. High frequency trading (HFT) programs were among the buyers: they *quickly bought and resold* contracts to each other, creating a **hot-potato** volume effect, and combinied sales drove the E-mini price down 3% in just 4 minutes.

Many other notable market crashes occurred since then, including on the day when the Associated Press (AP) Twitter account was *hacked* announcing that the White House had been bombed and President Obama injured,causing the Dow Jones Industrial average (DJIA) to loose 140 points (which were recovered in minutes). Note also that the NASDAQ was the victim of several mini flash crashes in 2012. ıindex[sub]mini flash crash

We close this introduction with a short list of the many topics related to high frequency trading which we shall not discuss in this introductory chapter, providing references for the interested reader to find information. For example, we shall not discuss price manipulation or arbitrage. The interested reader is referred to the fundamental work of Kyle and Viswanathan [66] or [2], [3], [9, 49], [51], [59, 58]

Neither will we consider asset pricing in the presence of liquidity constraints like in [21], [1], [35] Even though many *low frequency applications* require and/or benefit from optimal execution algorithms, the latter became an integral part of electronic trading in general, and high frequency trading in particular, and we shall concentrate on problems of optimal execution in an order book model like in [18], [22] or in a model trying to capture price impact in a diffusion model like in the early work of Bertsimas and Lo [19], Almgren and collaborators[10, 11, 13, 12, 14], and the fundamental work of Obizhaeva and Wang [79] followed by the series of works of Schied and collaborators [4, 5, 6, 7, 88, 89]. See also [85]. [23], [62] [55] However, we shall not dwell on models offering the several liquidity venues to choose from. to this effect, smart order routing was developed to help traders seek out where prices are best across a range of competing exchanges, platforms and dark pools, and route their orders accordingly for execution. Theoretical and practical optimization of execution in these conditions are discussed in [67], [42] or [64] for example.

Early discussions of the empirical properties of LOBs can be found in [82]. However, it is clear that the properties of the LOBs and their dynamics have been a favorite of the early econo-physics literature. Notable works include [17], [73, 24, 25, 26, 84], [47, 45], [57]

The interested reader is referred to the books of Hasbrouck [56], O'Hara [80], LeHalle's recent survey [69] or the recent book [68].

Dynamic models of LOBs are discussed in [41, 43] following early work of Rosu [86].

In the finance literature, most of the important contributions owe to the fundamental works of Glosten and Milgrom [52] on competitive equilibriums between informed and uninformed traders and Kyle and his collaborators [65] on the role of noisy traders.

Acknowledgements: I would like to thank Philip Oasis and Kevin Webster for sharing their passion for the high frequency markets, and for helping me understand the idiosyncrasies of the NASDAQ ITCH data.

1.2 Limit Order Books (LOBs)

Limit Order Books (LOBs) are today's financial markets response to price formation. By matching buyers and sellers on electronic platforms, they provide real-time access to traders all over the world. By understanding their dynamics, we gain insight in the role of limit and market order trading, and by abstracting their effects from the microscopic level to the macroscopic level, it becomes possible to propose models of price impact consistent with the microstructure of these electronic markets, and so doing, offering a mathematical framework in which the important problem of optimal trade execution can be formulated and tractable solutions can be derived.

LOBs are a form of double auction style trading (see for example [72, 92] for a discussion of this mechanism). Our presentation of the LOBs is rather simplistic. It is meant as an introduction in the style of book series *LOBs for Dummies* or introductory classes *LOBs 101*, as it should be for a *practitioner want-to-be*, with no ambition to raise to the level of the sophisticated models introduced by economists and econo-physicists, but with enough relevant information to understand the mathematical models which we introduce subsequently and benefit from their analyzes. We proceed to give some formal definitions needed to understand the process of limit order book trading.

In limit order markets, every market participant can post buy and sell orders in multiples of a minimal number of units of the financial interest being traded. This minimal number of units which can be traded is called the lot size. So when we talk about an order of size k, or a volume k, we mean an order for $k\ell$ units where ℓ denotes the lot size.

Definition 1. *A buy (resp. sell) limit order of size k at price p is a commitment to buy (resp. sell) up to k lots of the financial interest being traded at a price no greater than (resp. no less than) p.*

When a buy (resp. sell) limit order is first submitted, a matching algorithm checks whether it is possible to match it to previously submitted sell (resp. buy) orders. If so, the matching occurs immediately. If not, the newly submitted order becomes active, and it will remain active until matched to an incoming sell (resp. buy) order, or cancelled.

It is precisely the active orders in a market that make up the LOB.

Definition 2. *The LOB $L(t)$ is the set of all active orders in the market at time t.*

The depth available at price p and time t is the total amount (in lot size) being requested for purchase (resp. for sale) at price p in the LOB $L(t)$.

- Prices are multiple of the tick size
- In most markets, for a given price, orders are arranged in a **First-In-First-Out** (FIFO) stack
- At each time t
  - The **bid** price B_t is the price of the highest **waiting buy order**
  - The **ask** price A_t is the price of the lowest **waiting sell order**
- The state of the order book is modified by order book events arrivals, typically a) limit orders; b) market orders; c) cancelations.
- When the stock is traded on several exchanges, a *consolidated order book* is constructed by aggregation over all the visible trading venues.

1.2.1 The Role of a LOB

The limit order book plays a crucial role in high frequency finance. It is the only way to understand price formation and the source of frictions such as transaction costs. Recall that trading on a limit order book requires that

- ***Passive traders*** also called ***liquidity providers*** post trading intentions in the form of bids and offers;

- ***Aggressive traders*** also called ***liquidity takers*** execute certain orders, creating a form of adverse selection.

There is an extensive literature describing the empirical properties of the LOBs. We refer the interested reader to [20], [73], [84],[93], [77], [37, 38] for complements on the idiosyncrasies of the LOBs.

Figure 1.1 gives a typical graphical representation of such a limit order book. We use two different colors to distinguish the buy (green) and sell (blue) limit orders. Clearly the green buy orders are offered at a lower price level than the sell orders, so they form a histogram of their own to the left of the best bid. In this chapter, we plot at most the 10 highest levels of bids in the left part of the plot, and the at most 10 of the lowest levels of ask (in the right part of the plot). By construction, the best bid is smaller than the best ask, for if it was not the case, immediate matching would remedy the situation. Because color graphics are not always an option, the volumes of the bid orders on the left part of the limit order book are sometime made negative for a monochromatic plot to still be able to distinguish between bids and offers. We shall use this convention later in section 1.6 when we discuss mathematical models for limit order book dynamics.

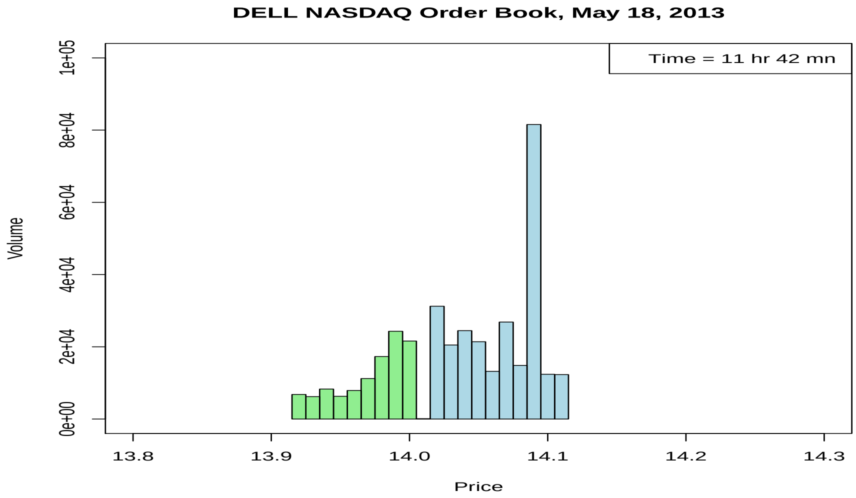

Fig. 1.1. DELL Limit Order Book on May 18, 2013

1.2.2 Limit Orders

A **limit order** sits in the order book until it is a) either executed against a **matching** market order; b) or it is **canceled**. It may be executed very quickly if it corresponds to a price near the bid and the ask, but it may take a long time to be executed if a) the market price moves away from the requested price; b) or the requested price is too far from the bid/ask. Most importantly, it is crucial to remember that it can be canceled at any time. Typically, a limit order waits for a match, but two things are certain: the cost of the transaction is **known** but the execution time is highly uncertain.

1.2.3 Market Orders

On the other end, a **market order** is an order to buy/sell a certain quantity of the asset at the **best available price** in the book. When an (impatient) agent puts a **market order**, say a buy (resp. sell) order: the first share(s) will be traded at the ask (resp. bid) price and the remaining one(s) will be traded some ticks higher (resp. lower) in order to fill the order size. The ask (resp. bid) price is then modified accordingly. whenever either the bid or ask queue is **depleted** by market orders or cancelations, and the price is **updated** up or down to the next level of the order book.

Typically a **market order** consumes the cheapest limit orders, but as before two things are certain: execution is **immediate** (at least if the book is deep enough) but the price per share is now **uncertain** as it depends upon the order size and the configuration of the order book.

1.2.4 Cancellations

Cancellations occur when the owner of an order no longer wishes to trade at the stated price, or when an order remained unmatched for too long and is automatically removed by the operator. Exchange specific rules have been designed to *clean up* the books of orders inactive for too long. When a program puts a cancellation of x lots in a given queue, the queue size is reduce by the amount x. When either the bid or the ask queue is depleted by market orders and/or cancelations, the *price* moves up or down to the next level of the order book.

The superior speed of some computerized trading programs offer the possibility to post orders for milliseconds with essentially no risk to be executed. This is an invitation to new forms of trading strategies, some of them at the very boundary of what can be considered as ethical, and now looked at seriously by the SEC. Quote stuffing is one of the many examples of these new trading strategies.

Quote stuffing is the practice of sending multiple orders in a short period of time with no real intention to transact, the orders being cancelled almost immediately. Order stuffing has been used to hide the originator's trading intentions, and in some cases to overload the system and slow down the competition by de facto withdrawing bandwidth. It can also be used to lure others into undesirable deals. See for example [81] for an example of the latter.

1.2.5 LOB Dynamics Summary

An agent aiming at a trade in a LOB market can proceed in two different ways:

- the agent can place a limit order and wait that the order is matched by another one, typically a crossing market order. In this case, the agent is said to be a passive trader, the cost of the transaction is known to the agent in advance, but the time it will take for the order to be executed is uncertain.
- Alternatively, the agent can place a market order that consumes the cheapest matching limit orders in the book. In this case, the agent is said to be an *aggressive* or *impatient* trader. He or she benefits from immediate execution, at least as long as the book has enough depth, but the price per share of the transaction is uncertain as it depends upon the order size and the current depth of the book.

To be more specific, for a buy (resp. sell) market order, the first lots/shares will be traded at the best ask (resp. best bid) price while the last ones can be traded several ticks upper (resp. lower) in order for the order to be filled in full. The best ask (resp. bid) price is then modified accordingly.

An agent can also place a **cancellation** order for a given number of lots/shares in a given queue, reducing the length of the queue by the corresponding amount. As we illustrate below, when either the best bid or ask queue is depleted by market orders and cancelations, the best bid or ask moves up or down to the next populated level in the order book, and the mid-price moves accordingly. These changes in the best bid or ask and the mid-price are the main source of transaction cost in LOB trading, and of adverse selection as the price moves in the direction of the market order.

1.2.6 Example of the Market Impact of a Large Fill

Let us assume that the current mid-price $p_{mid} = (p_{Bid} + p_{Ask})/2$ is equal to $p_{mid} = 13.98$, and that a buy of size $N = 76015$ arrives as a market order. We assume that the configuration of the LOB is such that only n_1 lots/shares are available at the best ask p_1, n_2 lots/shares are available at price $p_2 > p_1$, ..., and finally n_k shares are available at price $p_k > p_{k-1}$, making the order whole since $N = n_1 + n_2 + n_3 + \cdots + n_k$. The cost of the transaction is:

$$n_1 p_1 + n_2 p_2 + \cdots + n_k p_k = 1064578,$$

and the effective price is given by:

$$p_{eff} = \frac{1}{N}(n_1 p_1 + n_2 p_2 + \cdots + n_k p_k) = 14.00484.$$

Moreover, the new mid-price moved to $p_{mid} = 13.995$. The market impact of such a transaction is illustrated in Figure 1.3.

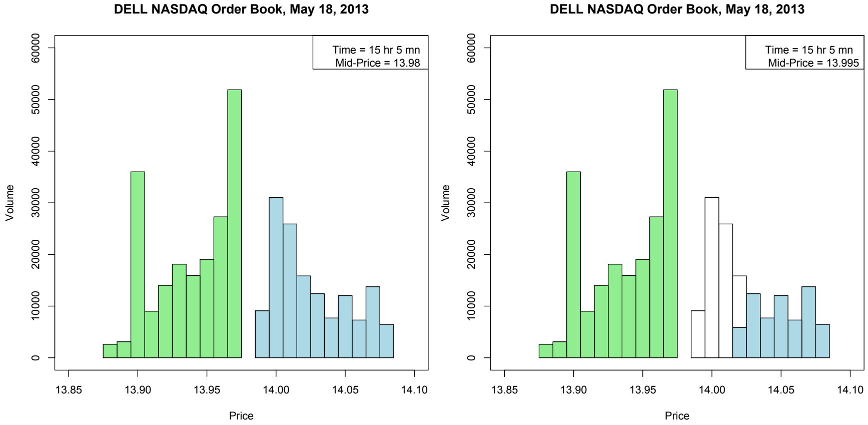

Fig. 1.2. Illustration of the market impact of a large fill as described in the text.

1.2.7 A LOB Idiosyncrasy: Hidden Liquidity

Some exchanges (e.g. NASDAQ & NYSE) allow hidden orders to be posted. The presence of these limit orders is revealed after their executions as they are made *visible* to the broader market after being executed. Allowing hidden orders is a common practice, but it is still very controversial as it acts as a barrier to the implementation of a fully transparent market place, and remains an impediment to price discovery and information dissemination. Empirical analyzes (see for example [78]) have proven that the presence of hidden liquidity encourages *fishing*: after it is revealed that a hidden order was executed, a significant increase of order placement inside the bid-ask has been documented. High frequency traders responsible for this rash of activity inside the bid-ask (whose identities are not known) can be divided in two groups: 1) Traders who try to take advantage of the remaining hidden liquidity; 2) Traders who try to steal execution priority from the fully hidden orders.

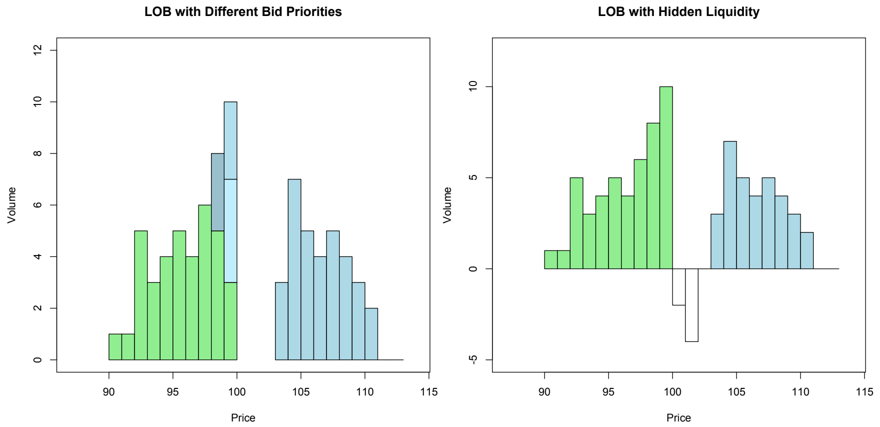

Fig. 1.3. Illustration (left) of one priority mechanism (if a sell market order comes in, the top part of the best bid bar will be executed, then the middle part, . . .) and of the possible presence of hidden liquidity (right).

1.2.8 ”Partially Hidden” Orders: Iceberg Orders

Most exchanges are not entirely lit, and allow *dark liquidity* to be posted inside the LOB and be partially visible only. These orders are called iceberg orders. They have two components: the *shown quantity* and the *hidden remainder* They are queued with the visible part of the LOB, and obviously, only the shown quantity is visible. When the order reaches the front of the queue, *only* the displayed quantity is filled. At that time, the trade (price & quantity filled) is *revealed* and the hidden part of the iceberg order is moved to the visible part of the book and placed at the back of the queue. Some exchanges charge extra execution fees for these types of order.

Iceberg orders are used by brokers (mostly electronic) who need to place large orders and try to avoid the negative effect of the price impact resulting from the large size (a form of adverse selection) and fear to be taken advantage of by predators. For this form of limit order to be rational, these benefits have to outweigh the loss of priority (and the lower likelihood to be executed), and possible extra fees.

1.2.9 Dark Pools / Crossing Networks

Dark pools are anonymous trading venues where quotes on securities or FX are not displayed publicly and trades are executed anonymously. They are run by private brokerage firms, e.g. Liquidnet, Pipeline, ITG’s Posit, Goldman’s SIGMA X. In the US, they are regulated by the SEC as Alternative Trading Systems. Their *raison d’être* is to move large numbers of lots without impacting the price or requiring iceberg orders. Typically, they consist of electronic engines that match buy and sell orders without routing them to lit exchanges. Participants submit (wish) lists of orders to a matching engine. and matched orders are executed at the midpoint of the bid-ask spread.

Dark pools appeal to traders who need large transactions. Indeed, the latter can rarely be executed without creating adverse price impacts. Their proponents argue that trading in a dark pool at the mid-point can be better than on a lit market On the other side of the divide, opponents complain about the lack of control and having to wait a long time for a match to occur.

Dark pools remain controversial because of the lack public disclosure, and despite the complains of many market participants, not much has been done to increase *transparency*. This is the more relevant that they represent a significant proportion of the trading activity. Indeed, it was estimated that trading on dark pools represented approximately 32% of the trades in 2012 (!).

This justifies the fact that part of the research on optimal execution includes models in which the broker has access to dark pools. Including such a possible alternative venue to route trades increases significantly the complexity of the optimization problem. The interested reader is referred to [64] for an example of a possible approach.

The reader interested in any of these forms of hidden liquidity is invited to consult the report [60] which, even though a few years old, is still quite informative on the subject.

1.3 Accounting Relationships on a Limit Order Book and Low Frequency Trading Models

The main purpose of this section is analyze the microstructure of the limit order book markets from a purely accounting perspective, trying to keep track of the impact of limit and market orders on the inventories, cash holding and wealth of the market participants. We use NASDAQ ITCH data which includes all limit and market orders and allow for a perfect reconstruction of the visible limit order book. For the sake of illustration, we consider the case of KO (Coca Cola) on 18/04/13. One of our goals is to differentiate between the agents using market orders (whom we also call liquidity takers or aggressive traders) and the agents using limit orders (whom we also call liquidity providers or passive traders).

In the second half of this section, we consider possible transitions from discrete to continuous models, trying to allow the latter (mostly diffusion models intended for low frequency traders) to incorporate macroscopic forms of the idiosyncrasies of the order books and the accounting relationships we identified at the microscopic level. We shall use the so-called *self-financing condition* as a testbed for our demonstration of how the low frequency trading models should incorporate the properties of high frequency markets, and we shall provide a simple example of option hedging to highlight the provocative consequences of the transition.

The material of this section is borrowed from [32, 31].

1.3.1 Evolution of the Limit Order Book in the Trade Clock

Following [81], we analyze the succession of trades in an event clock. So for us, time is discrete and $n = 1, ...N$ correspond to the times $t_1 < ... < t_N$ at which trades occur . For any quantity x_n depending upon the trade time n, we denote by $\Delta_n x = x_{n+1} - x_n$ the forward increment. We denote by p_n the mid-price just before the trade taking place at time t_n. So in some sense, we could think of p_n as of p_{t_n-} if these mid-prices were the samples of a continuous time p_t.

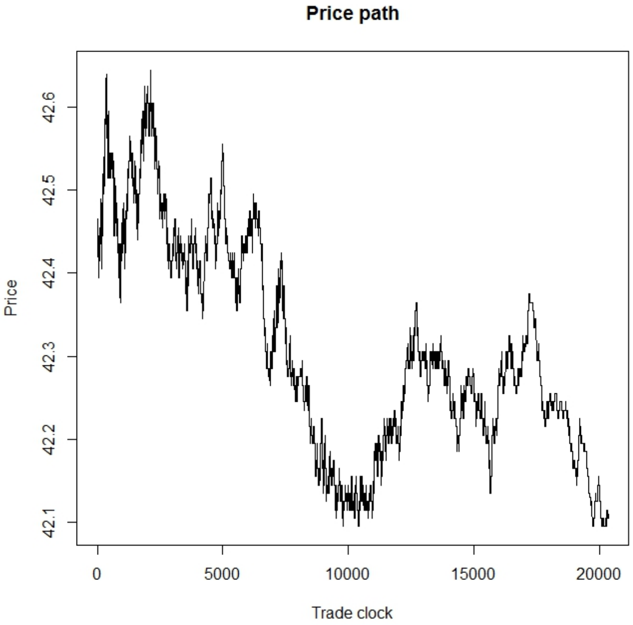

Fig. 1.4. Mid-price time evolution for KO on April 18, 2013 as a function of the trade clock time n

For the first part of this discussion, we assume that All the trades happen at the best bid or best ask. This means that for all practical purposes, the limit order book is nothing more than a stack of limit orders sitting on the best bid and another stack sitting on the best ask ! The fact that this appears to be overwhelmingly the case on NASDAQ ITCH data is merely an artifact of the interpretation of the successive messages comprising the ITCH data. We take them at face value for the time being, postponing to a later subsection the reconstruction of large orders from their *child orders* and the construction of a limit order book with depth on both sides.

We denote by s_n the bid-ask spread (i.e. the distance between the best ask and the best bid) just before the trade. Empirical studies have argued that $s_n \approx |\Delta_n p|$, claim which we shall revisit later on.

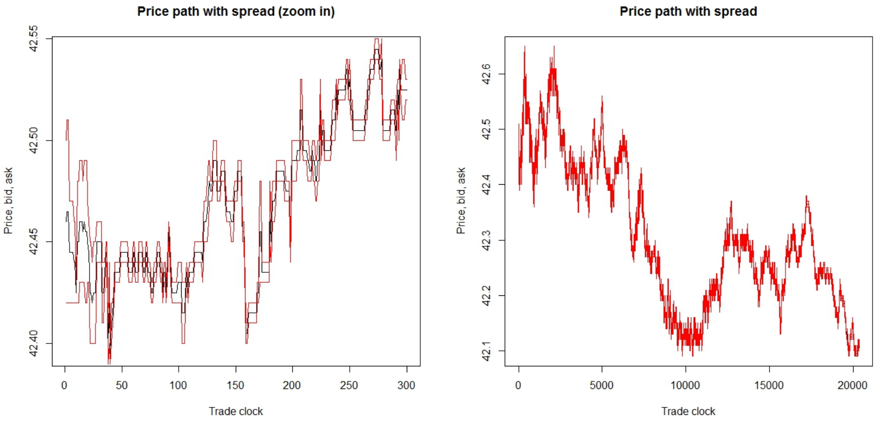

Fig. 1.5. Bid-ask spread time evolution for KO on April 18, 2013 as a function of the trade clock time n over a short period of the day (left) and over the whole day (right).

Most of the ITCH messages do not include the identity of the agent whose action triggered the message. So since we cannot keep track of *who does what*, for the sake of definiteness, we choose to work with the aggregate passive trader, and by duality, the aggregate aggressive trader. So we denote by L_n the inventory of the aggregate passive trader (liquidity provider) just before the n-th trade. In particular, $\Delta_n L < 0$ means that a market order bo

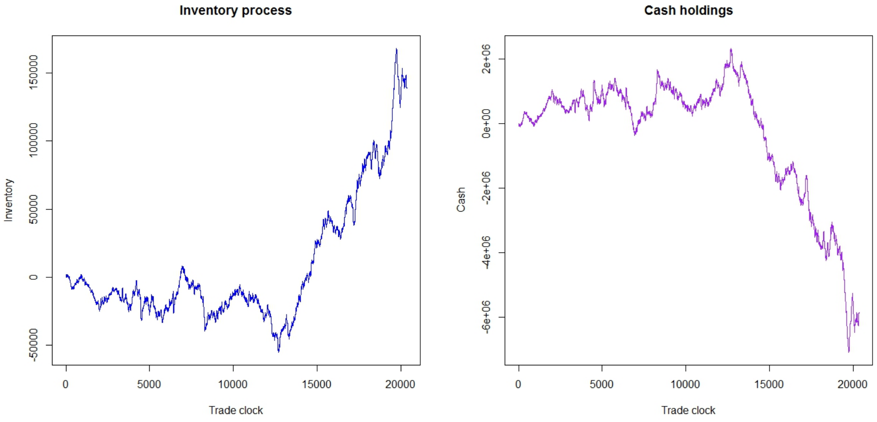

Fig. 1.6. Inventory (left) and cash account (right) of the aggregate liquidity provider for KO on April 18, 2013 as a function of the trade clock time n.

It is argued in [32] that $\varDelta_n L \varDelta_n p \leq 0$ holds 99.1% of the time and statistical tests are used to show the overwhelming significance of this hypothesis. In fact, its interpretation is quite intuitive: prices move in favor of market orders. This price impact of the market orders is a form of *adverse selection*.

Continuing our search for accounting relationships, we introduce K_n as the level of cash held by the aggregate passive trader just before the trade. By convention the changes in cash need to be the amounts exchanged during trades, no more, no less. This is a form of the self-financing property used frequently in financial mathematics.

Finally, we define the wealth of the aggregate passive trader (liquidity provide) as the quantity

$$X_n = L_n p_n + K_n$$

Our accounting convention is that the wealth is the value of the inventory *as marked to the mid-price* plus the cash holdings.

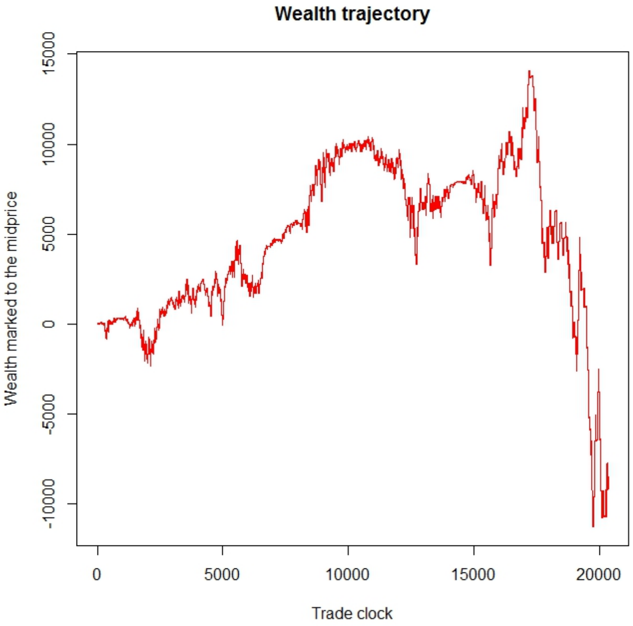

Fig. 1.7. Wealth of the aggregate liquidity provider for KO on April 18, 2013 as a function of the trade clock time n.

1.3.2 Self-financing equations

We consider three possible wealth dynamics:

$$\Delta_n X = L_n \Delta_n p \tag{1}$$

$$\Delta_n X = L_n \Delta_n p + \frac{s_n}{2}|\Delta_n L| \tag{2}$$

$$\Delta_n X = L_n \Delta_n p + \frac{s_n}{2}|\Delta_n L| + \Delta_n p \Delta_n L \tag{3}$$

These relationships can be rewritten in terms of the cash account values instead of the wealth. They read

$$\Delta_n K = -p_{n+1}\Delta_n L \tag{1.1}$$

$$\Delta_n K = -p_{n+1}\Delta_n L + \frac{s_n}{2}|\Delta_n L| \tag{1.2}$$

$$\Delta_n K = -p_n \Delta_n L + \frac{s_n}{2}|\Delta_n L| \tag{1.3}$$

In both equivalent formulations, the first equation is a form of the classical self-financing condition of the classical Black-Scholes theory. The second equation is reminiscent of studies on proportional transaction costs as the correction term involves the size of the transaction $|\Delta_n L|$ (irrespective of the direction buy or sell) multiplied by the spread. Finally, the third equation in the first set of bullet points includes a price impact term.

The obvious question is *which one is right?* We show that, when all the trades take place at the best bid or best ask, the third relationship holds true.

When a trade happens, the following five bullet points exhaust all the possibilities.

1. trader triggers a buy with a market order;
2. trader triggers a sell with a market order;
3. trader has his buy limit order executed;
4. trader has his sell limit order executed;
5. trader is not part of the current trade.

In case 1, the trader buys at the ask ($L_{n+1} - L_n > 0$)

$$K_{n+1} - K_n = -\left(p_n + s_n/2\right)\left(L_{n+1} - L_n\right)$$

In case 2, the trader sells at the bid ($L_{n+1} - L_n < 0$)

$$K_{n+1} - K_n = -\left(p_n - s_n/2\right)\left(L_{n+1} - L_n\right)$$

In case 3, the trader buys at the bid ($L_{n+1} - L_n > 0$)

$$K_{n+1} - K_n = -\left(p_n - s_n/2\right)\left(L_{n+1} - L_n\right)$$

In case 4, the trader sells at the ask ($L_{n+1} - L_n < 0$)

$$K_{n+1} - K_n = -\left(p_n + s_n/2\right)\left(L_{n+1} - L_n\right)$$

Finally, in case 5, $K_{n+1} = K_n$ and $L_{n+1} = L_n$.

The net result is that these five cases can be summarized into the single equation

$$\Delta_n K = -p_n \Delta_n L \pm \frac{s_n}{2} |\Delta_n L|$$

where ”±” means

- ”+” when trading with limit orders
- ”−” when trading with market orders.

Using the definition $X_n = L_n p_n + K_n$. of the wealth we get

$$\boxed{\Delta_n X = L_n \Delta_n p \pm \frac{s_n}{2} |\Delta_n L| + \Delta_n L \Delta_n p}$$

Figure 1.8 compares the aggregate wealths computed from the three self-financing condition *candidates* with the true wealth computed from the data, and already plotted in Figure 1.7.

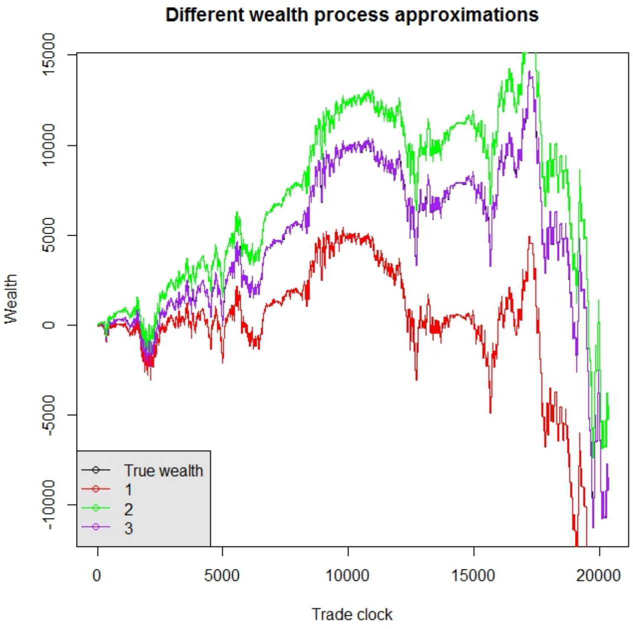

Fig. 1.8. Comparison of the aggregate wealths computed from the three potential self-financing condition with the true wealth.

The result is striking. It appears clearly that the true wealth coincides perfectly with the wealth computed with the third candidate. Moreover, we see that the difference between the wealths computed from the first two equations is large, proving the significance of transaction costs. Finally, it is clear from this plots that the difference between the last two self-financing equations, as given by the price impact term, cannot be neglected.

Prompted by this surprising graphical evidence, the natural questions should be: *Did We Miss Something?*, and *Could we be misled by artifacts of the special nature of NASDAQ messaging system?*

Despite the fact that all the ITCH messages overwhelmingly point to trades at best-bid or best-ask, we believe that this is the result of large order splitting into smaller (child) orders which can be executed against waiting limit orders from a single agent. So we developed and implemented simple algorithms reconstructing the *parent orders* from the *child orders*, though we can never be sure that we actually reversed engineered the NASDAQ order splitting process. Using these aggregate messages, we cannot assume that a typical order book is merely a combination of two *delta functions* as it is now the superposition of two full histograms. The good news is that, even with more general LOB models, we can still derive a self-financing condition with three similar terms playing the same roles, and test its validity on data from our reconstructed transactions. We refer the interested reader to [32, 31] for details and more applications. Here we limit ourselves to state the results without proofs.

1.3.3 The Case of More General Order Books

We assimilate the histograms forming the order book, after centering around the mid-price, to a non-negative continuous function and we define the order book shape function γ as the second anti-derivative of this function. So $\gamma''(u) \geq 0$ and $\gamma''(0) = 0$. γ is convex by definition. Next, we define the *transaction cost function* c as the Legendre transform of γ. More precisely:

$$c(\ell) = \sup_u \left(u\ell - \gamma(u)\right).$$

As a consequence, the function c is convex and satisfies $c(0) = 0$. The construction of the function γ from the order book,is illustrated in Figure 1.9. Notice that

- $c(\ell) = c\ell^2$ corresponds to **flat** order book
- $c(\ell) = c|\ell|$ corresponds to **best bid / best ask** order book

In the above setting of a general order book function, we can still derive the form of the self-financing condition. It is given by

$$\Delta K = -p\Delta L + c\left(-\Delta L\right), \tag{1.4}$$

in its cash form, and the wealth form generalizes to

$$\Delta X = L\Delta p + c(-\Delta L) + \Delta p\Delta L. \tag{1.5}$$

As earlier, we favor the form (1.5) of the self-financing condition because of the intuitive interpretation of its three terms: frictionless changes in portfolio value, transaction costs, and price impact / adverse selection.

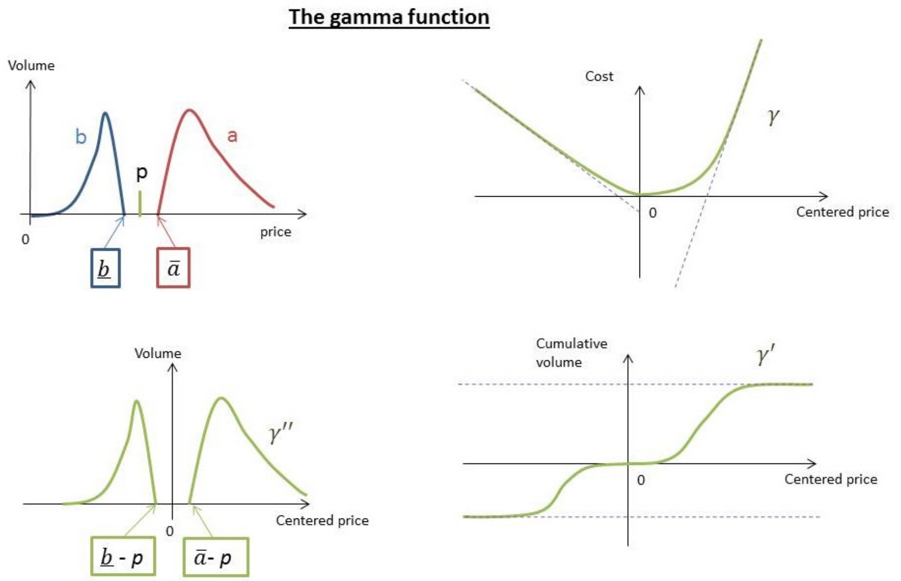

Fig. 1.9.

1.3.4 Continuous Limits and Slow Frequency Trading Models

We now consider a market model in which a quantity p_t, which we assimilate to a mid-price at time t, is given by an Itô's diffusion and a Low Frequency Trader (LFT) inventory L_t at time t which is also given by an Itô's diffusion. So we assume that:

$$\begin{cases} dp_t &= \mu_t dt + \sigma_t dW_t \\ dL_t &= b_t dt + \ell_t dW'_t \end{cases} \tag{1.6}$$

for some adapted processes μ_t, b_t, $\sigma_t > 0$, $\ell_t > 0$, which we may want to assumed to be bounded for the purpose of the present discussion, and two dependent Wiener processes W and W' satisfying

$$[W, W']_t = \int_0^t \rho_s ds$$

for some adapted process $\rho_t \in [-1, 1]$. Next, we assume that s_t is a continuous adapted process. This will be our proxy for what happened to the spread in the continuous time limit.

Our goal is to identify the forms taken in this continuous time set-up, by the self-financing conditions identified earlier at the level of trading on the order book. While the impact of the choice of a discretization is not innocent (see [32, 31] for a discussion of the different limits which can be obtained from different discretization choices), we shall choose

$$p_n^N = p_{\lfloor n/N \rfloor}; \quad L_n^N = L_{\lfloor n/N \rfloor}$$

and

$$s_n^N = \frac{1}{\sqrt{N}} s_{\lfloor n/N \rfloor}.$$

So if we believe that these discrete time processes correspond to the mid-price, the inventory of the aggregate passive trader, and the spread in a LOB market, then if we first consider the simpler case of trading exclusively at the best bid or best ask, the wealth dynamics ought to be given by the formula

$$\varDelta_n X^N = L_n^N \varDelta_n p^N + \frac{s_{\lfloor n/N \rfloor}}{2\sqrt{N}} |\varDelta_n L^N| + \varDelta_n p^N \varDelta_n L^N$$

and the price impact constraint

$$\varDelta_n p^N \varDelta_n L^N \leq 0$$

should hold if we want the discretization to *mimic* the micro structure of a LOB.

Because of the specific scaling we chose to obtain the discrete time processes from the continuous time processes we started from, we can use standard diffusion limit arguments to control the limit $N \to \infty$ and return to the initial low frequency set-up. The wealth processes of the discretized trading models converge to a limit X which naturally represents the wealth of the aggregate passive trader:

$$X_t = \lim_{N \to \infty} X_{\lfloor Nt \rfloor}^N$$

in u.c.p. and the dynamics of this wealth process are given by:

$$dX_t = L_t dp_t + \frac{s_t \ell_t}{\sqrt{2\pi}} dt + d[L,p]_t \tag{1.7}$$

in the case of the Best-Bid / Best-Ask order book trading model, and in the general case

$$dX_t = L_t dp_t + \varPhi_{l_t}(c_t) dt + d[L,p]_t \tag{1.8}$$

with

$$\varPhi_\sigma(F) = \int F(y) \phi_{\sigma^2}(y) dy$$

Notice that the price impact constraint, a necessary condition for an inventory obtained by limit orders becomes:

$$d[L,p]_t \leq 0. \tag{1.9}$$

Proofs and further discussions are given in [32, 31]

1.3.5 An Application to Option Hedging in a Complete Model

For the sake of this application, we assume that the midprice p_t is given exogenously by a regular diffusion process

$$dp_t = \mu(p_t)dt + \sigma(p_t)dW_t, \tag{1.10}$$

and we assume that the limit order book shape function γ_t is given by a continuous process in the form of a Markovian function of the price level p. More precisely we assume that it is of the form

$$\gamma_t(\alpha) = \gamma(p_t, \alpha),$$

for some deterministic function $(p, \alpha) \hookrightarrow \gamma(p, \alpha)$. In this application, the admissible inventories are Itô processes of the form

$$L_t = L_0 + \int_0^t b_u du + \int_0^t l_u dW_u \tag{1.11}$$

where l_t is *signed*. Typically,

- $l_t < 0$ when trading with limit orders
- $l_t \geq 0$ when trading with market orders.

We define the function

$$g(p, l) = \text{sign}(l)\Phi_l(c(p, \cdot)) \tag{1.12}$$

for later convenience. Notice that the model is complete since both diffusion processes p_t and L_t are driven by the same Wiener process. We denote by $f \in C^0$ the payoff function of a European option with maturity T, by K_0 the trader's initial cash endowment, and given the nature of the self-financing condition, the value at time t of the hedging port olio by

$$X_t = L_0 p_0 + K_0 + \int_0^t L_u dp_u + \int_0^t (\sigma(p_u) l_u - g(p_u, l_u))\, du + r \int_0^t (X_u - p_u L_u) du$$

Definition 3. *An initial cash endowment K_0 and an inventory process L_t replicate the European payoff $f(p_T)$ at maturity T if*

$$X_T = f(p_T) \qquad \textit{and} \qquad X_0 = K_0 + p_0 L_0. \tag{1.13}$$

NB: When quoting a price for the option, the trader needs to quote an initial delta asked of the buyer of the option !

The following proposition is the mere result of a classical verification argument applied to the present situation:

Proposition 1. *If $f \in C^0$, $T > 0$ and $v \in C^{1,3}$ satisfies:*

$$\frac{\partial v}{\partial t}(t, p) + g\left(p, \sigma(p)\frac{\partial^2 v}{\partial p^2}(t, p)\right) - \frac{\sigma^2(p)}{2}\frac{\partial^2 v}{\partial p^2}(t, p) + rp\frac{\partial v}{\partial p}(t, p) = rv(t, p)$$

with terminal condition $v(T,p) = f(p)$, then

$$L_t = \frac{\partial v}{\partial p}(t,p_t), \quad \textit{and} \quad K_0 = v(0,p_0) - \frac{\partial v}{\partial p}(0,p_0)p_0$$

replicate the payoff $f(p_T)$ at maturity T, its volatility is given by

$$l_t = \sigma(p_t)\frac{\partial^2 v}{\partial p^2}(t,p_t),$$

and the replication price of the option is $X_0 = v(0,p_0)$.

The main *take-home message* from the present model is the following striking conclusion:

Corollary 1. *In the above complete model:*

- *Positive gamma options can only be hedged with market orders !*
- *Negative gamma options can only be hedged with limit orders !*

For the sake of illustration, we go over the argument and the conclusions in the special case of Best-Bid / Best-Ask limit order books. Under the same model assumptions (1.10) and (1.11) as above, we state a continuous time limit for the spread by assuming

$$s_t = \sqrt{2\pi}\lambda\sigma_t, \qquad \text{with} \quad \lambda > 1/2$$

This relationship between the spread and the mid-price volatility has been argued in the empirical literature, and it is rather universally accepted. Notice that choosing $\lambda = 1$ implies that the self-financing condition becomes $dX_t = L_t dp_t$ since the last two terms cancel each other, reducing this condition to the frictionless case! For the sake of simplicity, we assume that interest rates and dividends are null

We now address the following question: Given the model for p and s, find L such that X hedges a European option with payoff $f(p_T)$. Working in the Markovian set-up, we posit that the price of the option at time t is given by a function $v(t,p)$, if the mid-price is p, and assuming that this function is smooth, we apply Itô's formula and get:

$$\begin{aligned} d(X_t - v(t,p_t)) = (L_t - \varDelta_t)\, dp_t - (\varTheta_t + \frac{1}{2}\varGamma_t\sigma^2(t,p_t))dt \\ + \frac{s_t}{\sqrt{2\pi}}\ell_t dt + d[p,L]_t \end{aligned}$$

where we isolated on the second line the extra term due to the special form of the self-financing condition we are using, from which we obtain (by mere matching of the two possible Itô decompositions)

$$L_t = \varDelta_t$$

which also implies

$$-\ell_t = \Gamma_t \sigma(t, p_t)$$

So the final solution of the Delta hedging procedure is given by

$$\begin{cases} L_t & = \Delta_t \\ \ell_t & = -\Gamma_t \sigma(t, p_t) \end{cases}$$

and we recover the conclusion highlighted earlier: only negative Gamma options can be replicated via limit orders! As a final remark, we notice that the pricing PDE is given by

$$\partial_t v(t,p) + \left(\lambda - \frac{1}{2}\right) \sigma^2(t,p) \partial_p^2 v(t,p) = 0$$

which is the same type of PDE as the pricing PDE from the Black-Scholes theory, except for the fact that the local volatility is now multiplied by a factor of $\sqrt{2\lambda - 1}$

1.4 Heterogenous Beliefs and High Frequency Market Making

The goal of this section is to discuss market making in the context of high frequency trading. Early approaches based on agent based models can be found in Hasbrouck's book [56] and in the surveys [37, 38] by Chakrborti, Toke, Patriarca and Abergel. Inventory models were used by Garman in [50], and Amihud and Mendelson in [15]. The most famous model among economists is without a shadow of a doubt Kyle's model of informed traders [65]. See also the book of Maureen O'Hara [80]. Zero-intelligence models have been used by Gode and Sunder in [53], Maslov and Mills in [73], and Cont in [40]. Finally, we mention the models of market impact which are the most popular among practitioners as well as financial mathematicians. See for example Almgren-Chriss [13, 12], Bouchaud-Potters [84] or Schied [87].

The following discussion is based on [30]. The goal is to set up a tractable stochastic agent-based model in which realistic properties of the limit order book appear endogenously as outputs of the analysis, and not as hypotheses. It is important to emphasize that we are not searching for optimal execution strategies. We are merely proposing and analyzing a model for a possible interaction between a representative market maker and his clients.

1.4.1 The Agents

We first describe the market participants.

- One market maker.
  According to the definition given by NASDAQ, this is an agent who places competitive orders on both sides of the order book in exchange for privileges. For the purpose of the present discussion, we shall assimilate this agent to a liquidity provider, someone who posts an order book or equivalently, a transaction cost curve. As we shall see, his trategy will be to adjust pricing and volumes by guessing the client order flows.

- Clients.
  For the purpose of this discussion, we shall identify the clients to liquidity takers, agents who trade with the market maker. These clients place market orders, each client acting according to his/her *own information*.

The market maker model is based on the assumption that the clients are rational. He optimizes his order book choice. On the other end, given the order book posted by the market maker, the clients use their own information to capture he dependence between trades and price dynamics.

1.4.2 Mathematical Setup

We first describe the anatomy of a trade in our market model. We assume that the midprice P_t is given exogenously and is announced by the market at time t. Once this knowledge is made public, the market maker proposes an *order book* around P_t, namely a distribution of bids below the mid-price and a distribution of asks above the mid price. It is understood that the market maker cannot differentiate clients *pre-trade*. Then, each client can trigger trades. In order to obtain a volume l_t of the stock, a client needs to pay $P_t l_t + c_t(l_t)$, the function $\ell \hookrightarrow c_t(\ell$ representing the transaction cost function at time t. Finally, the identity of the client is revealed to the market maker *post-trade*. While this last assumption may seem inaccurate depending upon the market, it is generally true for FX trading.

It is important to emphasize that the market maker controls the transaction costs via his choice of the transaction cost function $\ell \hookrightarrow c_t(\ell)$. while for each index i, client i controls his trading volume l^i_t. We assume that

1. The marginal costs are defined in the sense that the function $\ell \hookrightarrow c_t(\ell)$ is differentiable in ℓ;
2. The clients may choose *not to trade*, i.e. $c_t(0) = 0$;
3. The midprice is well defined in the sense that $c'_t(0) = 0$;
4. The marginal costs increase with volume in the sense that c_t is convex;
5. c_t is equal to ∞ outside an interval.

The duality relationship between order book distribution function γ_t and transaction cost function c_t is crucial to the following developments. For each time t, the function $\alpha \hookrightarrow \gamma_t(alpha)$ is defined as the Legendre transform of the transaction cost function c_t:

$$\gamma_t(\alpha) := \sup_{l \in \text{supp}(c_t)} (\alpha l - c_t(l))$$

The fact that c_t is convex with compact domain implies that γ''_t is a positive finite measure. This distribution γ''_t represents the order book formed by the limit orders of the market maker. If γ''_t has a density $f(x)$, it is given by the *shape function* which we used earlier.

Mathematically speaking, the assumption of heterogeneous beliefs means that each agent (the market maker maker as well as each individual client) has his own information represented by a specific filtration, and his own probability measure on

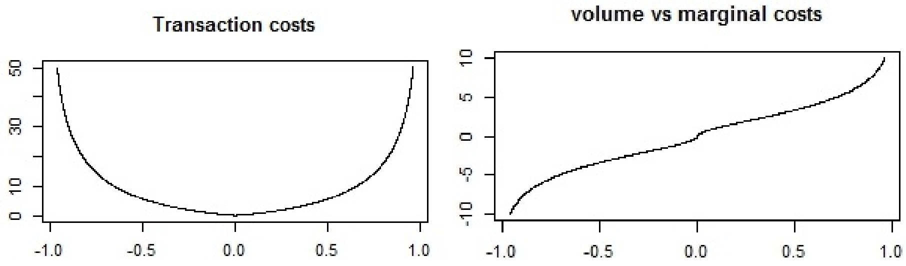

Fig. 1.10. Transaction cost function $l \hookrightarrow c_t(l)$ (left) and its derivative (right) giving the virtual market impact.

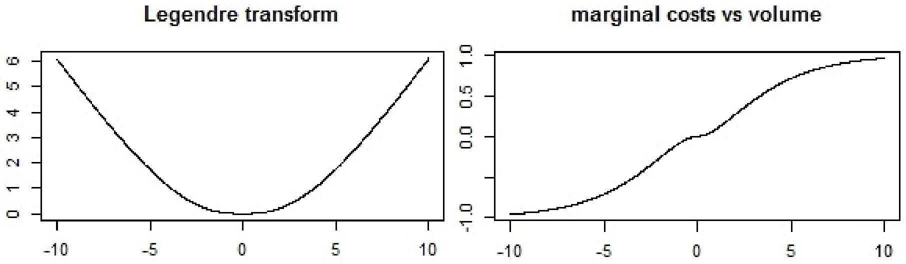

Fig. 1.11. Plot of the graph of the Legendre transform γ_t of the transaction cost function c_t (left) and its derivative (right) giving the marginal costs as functions of volume.

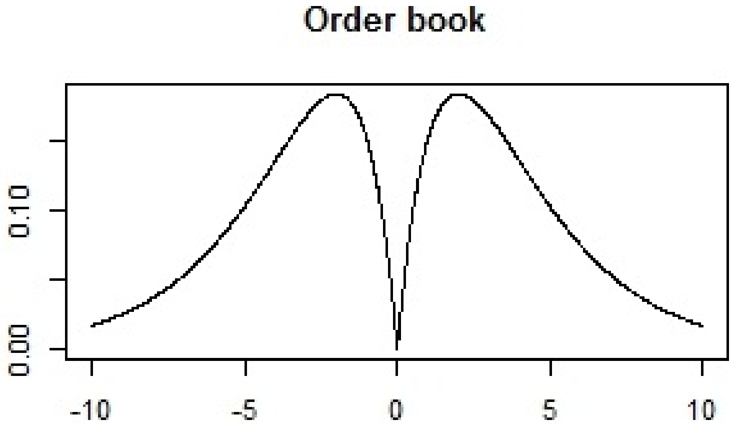

Fig. 1.12. Plot of the graph of the order book distribution $\gamma_t''(l)$.

this filtration. These filtrations (information structures) are potentially different, but we assume that the *price process is adapted to all of them* (i.e each client sees the price). So if we label the agents by $k = 0, 1, \cdots, n$, $k = 0$ standing for the market maker and $k = i = 1, \cdots, n$ for client i

1. $(\Omega, \mathcal{F}, \mathbb{F} = (\mathcal{F}_t)_{t \geq 0}, \mathbb{P})$ with W a $\mathbb{P}$-BM that generates $\mathbb{F}$.
2. $\mathbb{F}^k \subset \mathbb{F}$ generated by a $\mathbb{P}$-BM W^k.
3. $\mathbb{P}^k$ s.t. $\mathbb{P}^k|_{\mathcal{F}_t^k} \sim \mathbb{P}|_{\mathcal{F}_t^k}$.
4. P_t an Itô process adapted to *all* $\left(\mathbb{F}^k\right)_{k=0\ldots n}$.

1.4.3 Client Optimization Problem

As explained earlier, we assume that at each time t, the market maker *moves first* by posting his choice for the order book γ_t'' centered around the mid-price P_t which is common knowledge. Equivalently, he announces his choice for the transaction cost function c_t. Subsequently, the clients try to predict, as opposed to cause, the next price movements. In any case, their decisions do not affect c_t. The optimization problem for client i can be summarized as follows:

- *Exogeneous state variables*
  - P_t non-negative Itô process
  - c_t (random adapted) convex function in a fixed domain
- *Endogeneous state variables*

$$\begin{cases} dL_t^i = l_t^i dt \\ dX_t^i = L_t^i dP_t - c_t(l_t^i)dt \end{cases}$$

  Here, l_t^i appears as the control variable of client i, it is the rate at which he trades. L_t^i is his inventory, in other words the volume or total position of the client, and X_t^i is his wealth marked to the mid-price.
- *Objective function*

$$J^i = \mathrm{e}_{\mathbb{P}^i}\left[U^i(X^i_{\tau^i}, P_{\tau^i})\right]$$

  where U^i is the utility function of client i, and τ^i stopping time.

It is possible to solve this optimization problem and identify the optimal trading strategy.

Theorem 1. *Under suitable integrability assumptions on U^i and τ^i, the optimal strategy is given by:*

$$\alpha_t^i := c_t'(l_t^i) = \mathbb{E}_{\mathbb{Q}^i}\left[P_{\tau^i} - P_t | \mathcal{F}_t^i\right]$$

with $\frac{d\mathbb{Q}^i}{d\mathbb{P}^i} = \frac{\partial_X U^i(X^i_{\tau^i}, P_{\tau^i})}{\mathbb{E}_{\mathbb{P}^i}\left[\partial_X U^i(X^i_{\tau^i}, P_{\tau^i})\right]}$.

Remark 1. Testing the Client Model It is possible to test specific forms of the client model empirically. For example, if we assume that

- Under $\mathbb{Q}^i$, the stopping times are exponential, say $\tau^i \sim \exp\left(\beta^i\right)$ and independent of P_t.;
- The volatilities are controlled by the spread in the sense that

$$\sigma_t^i := | \underbrace{c_t'(l_t^i)}_{\textit{Implied alpha}} - \underbrace{(p_{\tau^i} - P_t)}_{\textit{Realized alpha}} | \leq \frac{\textit{spread}}{2},$$

leading to a *two parameter* model linking trade to price dynamics: (β^i, σ^i). If we further assume that all the clients have one of *two* time scales, and choose (β_1, β_2) that minimizes error between implied and realized alpha, we can use NASDAQ 'fullview' data (providing all public quotes and all trades with nanosecond timestamps) to test this form of the two time scale model. We refer to [31] for further details.

1.4.4 The Market Maker Optimization Problem

The market maker does not control his inventory. It is dictated by the trading decisions of his clients, so using primal variables, the dynamics of his inventory and his wealth are governed by the stochastic differential equations:

$$\begin{cases} dL_t = -\frac{1}{n}\sum_i l_t^i dt \\ dX_t = L_t dP_t + \frac{1}{n}\sum_i c_t(l_t^i) dt \end{cases}$$

We use the transaction cost function c_t as control for the market maker, and assuming that the clients acts rationally, he can assume that $\alpha_t^i = c_t'(l_t^i)$ so equivalently $l_t^i = [c_t']^{-1}(\alpha_t^i) = \gamma_t'(\alpha_t^i)$. So equivalently, we can rewrite his optimization problem in terms of his dual variables as:

$$\begin{cases} dL_t = -\frac{1}{n}\sum_i \gamma_t'\left(\alpha_t^i\right) dt \\ dX_t = L_t dP_t + \frac{1}{n}\sum_i \left[\alpha_t^i \gamma_t'\left(\alpha_t^i\right) - \gamma_t\left(\alpha_t^i\right)\right] dt \end{cases}$$

Despite our simplifying assumption that the market maker is *risk-neutral* this form of Stackelberg game is difficult to solve in its full generality. So for the sake of illustration, we propose a computation of an optimal choice for the market maker if we remove the feedback coupling between the clients alphas and the choice of the market maker, and we model the dynamics of the clients' alphas exogenously. This reduces the game to an optimal control problem for the market maker, making his optimization problem tractable.

1.4.5 Approximation and Corresponding Solution

We model the dynamics of the clients α_t^i by the following system of coupled stochastic differential equations:

$$d\alpha_t^i = -\rho\alpha_t^i dt + \sigma dB_t^i + \nu dB_t^0$$

for some positive constant $\rho > 0$. The mean reversion nature of the drift corresponds to a form of *decay of information*. Here the Wiener processes $(B_t^k)_{t\geq 0}$ are assumed to be independent of each other for $k = 0, 1, \ldots$. The term dB_t^0 can be interpreted as a common noise coupling these dynamics. If we see these dynamics as a microscopic description of the model, we could wonder what would happen to the statistical distribution of the α_t^i when the number n of clients grows without bounds. This leads to a macroscopic description of the model. It is easy to see that the empirical distribution of the α_t^i for $i = 1, \cdots, n$ converges to a probability measure μ_t when $n \to \infty$. This limiting measure $\mu_t(d\alpha)$ can be interpreted as the client belief distribution. It is easy to see that the dynamics of these probability measures are given by the following stochastic partial differential equation:

$$d\mu_t(\alpha) = \left[\frac{1}{2}\left(\sigma^2 + \nu^2\right) \Delta\mu_t(\alpha) + \rho\nabla\left(\alpha\mu_t(\alpha)\right)\right] dt - \nu\nabla\mu_t(\alpha) dB_t$$

Even though we mentioned earlier that the mid-price process (P_t) was given exogenously, we refrain from making explicit assumptions for this price process because of the heavy set of assumptions we just made on the dynamics of the α_t^i. Instead, we would like to *infer* properties of the price from the client trades. Recalling the implied alpha relationship

$$\alpha_t^i := c_t'(l_t^i) = \mathbb{E}_{\mathbb{Q}^i}\left[\int_t^\infty e^{-\beta^i(t-s)} dP_s \middle| \mathcal{F}_t^i\right]$$

we introduce a *price proxy*

$$dP_t^\lambda := \sum_{i=1}^n \lambda^i \left(\beta^i \alpha_t^i dt - d\alpha_t^i\right)$$

for any set of non-negative weights λ^i summing up to one, i.e. for which $\sum \lambda^i = 1$. Choosing this set of weights appropriately, one can control of the expected quadratic error in prices

$$\mathbb{E}\left|P_t - P_t^\lambda\right|^2 \leq \epsilon^2 \frac{1}{n}\sum_i I(\mathbb{Q}^i, \mathbb{P}) \approx -\epsilon^2 \int_0^t \left\langle \log\left(\frac{\gamma_s''}{\mu_s}\right), \mu_s \right\rangle ds$$

where we used the standard notation I for the Kullback - Leibler distance between probability distributions given by their *relative entropy*, and where:

$$\epsilon = \sqrt{\frac{n}{\sum_i \left(\sigma^i\right)^{-2}}} \leq \frac{1}{n}\sum_i \sigma^i$$

Based on these, we formulate a mean field approximate control problem for the market maker interacting with the continuum limit of clients as given by the limit of their empirical distribution. The state dynamics are given by:

$$\begin{cases} dL_t &= -\left\langle \gamma_t', \mu_t \right\rangle dt \\ d\mu_t(\alpha) &= \left[\frac{1}{2}\left(\sigma^2 + \nu^2\right) \Delta\mu_t(\alpha) + \rho\nabla\left(\alpha\mu_t(\alpha)\right)\right] dt - \nu\nabla\mu_t(\alpha) dB_t \end{cases}$$

and the objective function to maximize is now:

$$J^{\lambda} = \int_0^{\infty} e^{-\beta t}\mathbb{E}\left[L_t \left\langle id, (\beta\lambda)_t \right\rangle + \left\langle -L_t \beta id + (id - \bar{\alpha}_t)\,\gamma_t' - \gamma_t, \mu_t \right\rangle\right] dt$$

under the constraint $\int_0^{\infty} \left\langle e^{-\beta t} \log\left(\frac{\gamma_t''}{\mu_t}\right), \mu_t \right\rangle dt \leq C$. Recall that we assume that the market maker is risk neutral. Despite the infinite dimensional nature of the state dynamics, we can solve this problem using an appropriate version of the Pontryagin stochastic maximum principle. We can formally write the Hamiltonian of the problem and solve the adjoint backward stochastic differential equation, leading to the market maker's 'shadow alpha:

$$\alpha_t^* = \left\langle id, \lambda_t + \frac{(\beta\lambda)_t - \beta\mu_t}{\beta + \rho} \right\rangle$$

and the maximized Hamiltonian:

$$\mathcal{H}(\gamma, \mu, \alpha^*) = \langle (id - \alpha^*)\gamma' - \gamma + \epsilon \log \gamma'', \mu \rangle$$

Introducing the profitability function:

$$m(\alpha) = \underbrace{(\alpha - \alpha^*)}_{\textit{spread}} \cdot \underbrace{\int_{\alpha}^{\infty} \mu}_{\textit{filling probability}} \quad \textit{if}\, \alpha \geq 0$$

the maximized Hamiltonian can be rewritten as:

$$\mathcal{H}(\gamma, \mu, \alpha^*) = \langle \gamma'', m \rangle + \epsilon \langle \log \gamma'', \mu \rangle$$

and the optimal strategy can be found as the solution of the equation:

$$\frac{\gamma''(\alpha)}{\mu(\alpha)} = \frac{\epsilon}{C - m(\alpha)}$$

where C is a renormalization constant. Figure 1.13 gives a sample example of optimal order book posted by the market maker in this model. Details can be found in [31]

1.4.6 Comments and Possible Extensions.

- In the above discussion, the mid-price (P_t) was given exogenously. However, minor modifications can extend the model to achieve, like in Kyle's model [65], a price process derived endogenously from a known (though random) terminal value P_T.
- In our model, the clients and the market maker are clearly separated. This set-up is consistent with some of the FX markets. However, in most equity markets, agents act both as liquidity providers and liquidity takers. This suggests a desirable extension of the model.

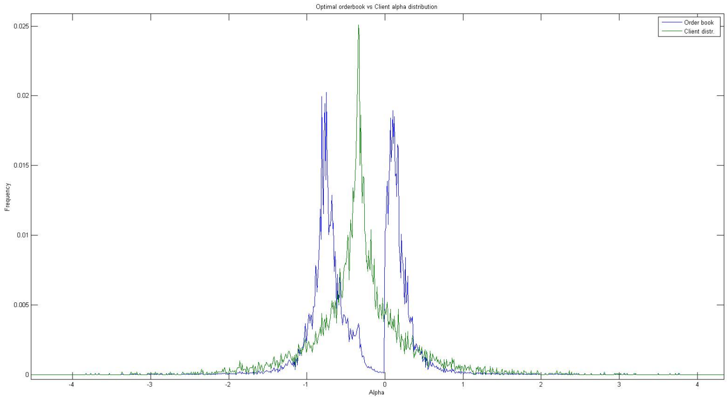

Fig. 1.13. Blue: Optimal order book γ''. Green: Client alpha distribution μ.

- We could postulate a nonlinear microscopic dynamics for the implied alphas of the clients in the form:

$$d\alpha_t^i = f(\alpha_t^i, \frac{1}{n}\sum_j \delta_{\alpha_t^j})dt + \sigma dB_t^i + \nu dB_t$$

  This would lead, at the macroscopic level, to a stochastic partial differential equation of the McKean-Vlasov type

$$\begin{aligned} d\mu_t = &\left[\frac{1}{2}\left(\sigma^2+\nu^2\right)\Delta\mu_t(\alpha) + \nabla\left(f(\alpha,\mu_t)\right)\mu_t(\alpha)\right]dt \\ &+ \nu\nabla\mu_t(\alpha)dB_t \end{aligned}$$

  for the distribution of the alphas, another challenging extension of our model.
- Order books are *additive* in the sense that, when there are multiple liquidity providers, one has:

$$\gamma''_{aggregate} = \sum_j \gamma''_j$$

  This suggests that extending the analysis to the case of multiple liquidity providers should be possible.
- Finally, we mention the challenging problem of the modeling of price impact from option trading. Clearly, trades contain information. Indeed, as we argued throughout, a trade is followed by a price movement, and the trade says something about the direction the traded asset is heading. However, if the traded asset is a *derivative*, the information about the underlying contained in the trade is now *non-directional* .

1.5 Predatory Trading

Instances of predatory trading are not limited to high frequency trading. As for the optimal execution problems discussed in Section 1.7, the models of predatory trading are used and implemented in applications with different time scales. However, their impact has been publicized (and often criticized) as part of the case against program trading in general, and high frequency trading in particular. So it is important to understand the reasons in the context of the mathematics of high frequency trading which is often assimilated to algorithmic trading.

1.5.1 Premises for Predatory Trading

The typical instances we should have in mind involve a large trader facing a forced liquidation, especially if the need to liquidate is known to other traders. They include hedge funds nearing margin calls, traders who use portfolio insurance, stop loss orders, etc., institutions and funds which cannot hold on to downgraded instruments, index tracking funds at re-balancing dates (e.g. when the next composition of the Russell 3000 is announced). These forced liquidations can be very costly because of significant price impact. Important examples to keep in mind include the famous Metallgesellschaft (MG) blunder, and the spectacular failure of the hedge fund Amaranth precipitated by the aggressive natural gas trades of his star trader Brian Hunter.

The following examples will be used as illustrations. The quotes highlight the issues we will emphasize in our discussions of the results of the mathematical models:

- In the case of the 1987 crash, the Brady's report states: ”several *triggers* ... ignited mechanical, price-insensitive selling by a number of institutions following portfolio insurance strategies ... The selling by these investors, and the prospect of further selling by them, encouraged a number of aggressive trading-oriented institutions to sell in anticipation of further declines.”
- In the case of the UBS Warburg's take over of Enron trading platform, ”UBS Warburgs proposal to take over Enrons traders without taking over the trading book was opposed on the ground that it would present a predatory trading risk, as Enron traders effectively knew the contents of the trading book.”
- In reference to the collapses of Long-Term Capital Management (LTCM) one can read in Business Week on February 26, 2001, ”... if lenders know that a hedge fund needs to sell something quickly, they will sell the same asset, driving the price down even faster. Goldman Sachs and other counter-parties to LTCM did exactly that in 1998. Goldman admits it was a seller but says it acted honorably and had no confidential information.”

As a final quote, we cite Jim Cramer, the host of NBC show Mad Money who once said, ” ... *When you smell blood in the water, you become a shark . . . . when you know that one of your number is in trouble . . . you try to figure out what he owns and you start shorting those stocks . . .*”

As illustrated by the above examples and quotes, most often, predatory trading takes the form of *front running*, whose instances can be illegal when triggered by inside information. In any case, we shall refrain from discussing the moral and ethical issues marring this practice and concentrate on mathematical models trying to capture the main stylized facts of predatory trading.

The review of the present section is based on the fundamental works of Brunnermeier and Pedersen [28], Carlin, Lobo and Viswanathan [29], and of the extension by Schied and Schöneborn [88] of [29].

1.5.2 Typical Predatory Trading Scenario

The goal of our theoretical analysis is to provide a set of mathematical models which can account for some of the most typical behaviors associated to predatory trading. Among them, we would like for example to account for the behavior of a distressed trader (whom we shall also call the prey) who needs to unload a large position in the presence of predators when trade size has a significant impact on price. One of the behaviors expected from the predators would be to initially trade in the same direction as the prey in order to 1) withdraw liquidity; 2) increase the market impact of the liquidation; 3) exacerbate the price change and force a costly over-shooting.

Next, we would like to see the model predict that the predator 1) reverses trading direction, profiting from the price over-shoot; 2) closes his position for a profit.

1.5.3 Multi-Player Game Model Mathematical Analysis

We now provide the details of the mathematical model. We assume that trading takes place in continuous time, in the presence of one risk free asset and one risky asset, that interest rate is zero, and that the market comprises $n+1$ strategic players and a number of noise traders. We denote by $X_0(t), X_1(t), \cdots, X_n(t)$ the risky asset positions of the strategic players. Working directly with a limit order book model would be too difficult for an equilibrium problem to be solved, so we choose to use an effective diffusion model based on the Almgren - Chriss linear price impact model. According to this model trades at time t are executed at the price

$$P(t) = \tilde{P}(t) + \gamma \sum_{i=0}^{n} [X_i(t) - X_i(0)] + \lambda \sum_{i=0}^{n} \dot{X}_i(t)$$

where $\tilde{P}(t)$ is a mean zero martingale, which we shall choose to be a Wiener process.

Our goal is to understand predation, illustrate the benefits of stealth trading, and sunshine trading in extreme market conditions such as 1) elastic markets for which the temporary price impact λ is much larger than the permanent price impact γ; 2) plastic markets in which the permanent price impact γ is much larger than the temporary price impact λ.

1.5.4 The One Period Game

We first concentrate on the simple case of one period models. We assume that each strategic player $i \in \{0, 1, \cdots, n\}$ knows the initial asset positions $X_j(0)$ for $j \neq i$

of all other strategic players as well as their target $X_j(T)$ at some fixed time point $T > 0$ in the future. We also assume that all the players are risk neutral in the sense that they seek to maximize their expected terminal wealth. In order to do so, they choose a trading strategy $X_i(t)$ satisfying their constraints $X_i(0)$ and $X_i(T)$ with the hope that it will end up being optimal. The set of players comprises one *distressed trader* whom we shall call *prey* from time to time and whom we cast as player 0.

$$X_0(0) = x_0 > 0, \qquad X_0(T) = 0,$$

and n *predators*, namely players $1, 2, \cdots, n$ for whom:

$$X_i(0) = X_i(T) = 0, \qquad i = 1, \cdots, n.$$

We now define the optimization problem leading (hopefully) to equilibriums. A strategy $X_i = (X_i(t))_{0 \le t \le T}$ is said to be *admissible* for player i if it is an adapted process with continuously differentiable sample paths. This last assumption is very restrictive, despite the fact that it is extremely popular with all the optimal execution problems based on price impact models. Indeed, the optimal solution of the Merton problem with proportional transaction costs does not satisfy this property as it has jumps (see for example [91]) and our earlier discussion of the self-financing condition for the high frequency markets argued that it could rule out price impact due to adverse selection.

In any case, given a set $\underline{X} = (X_0, X_1, \cdots, X_n)$ of admissible strategies each player $i \in \{0, 1, \cdots, n\}$ tries to maximize his expected return

$$J^i(\underline{X}) = \mathbb{E}[\int_0^T (-\dot{X}_i(t))P(t)dt]$$

under the constraint

$$P(t) = \tilde{P}(t) + \gamma \sum_{i=0}^{n} [X_i(t) - X_i(0)] + \lambda \sum_{i=0}^{n} \dot{X}_i(t).$$

Proving existence of *Nash equilibriums* and identifying their properties is the main mathematical challenge.

If we restrict the admissible strategies $\underline{X} = (X_0, X_1, \cdots, X_n)$ to be deterministic, then the objective functions become:

$$J^i(\underline{X}) = \mathbb{E}[\int_0^T (-\dot{X}_i(t))P(t)dt] = \int_0^T (-\dot{X}_i(t))\overline{P}(t)dt$$

where

$$\overline{P}(t) = P(0) + \gamma \sum_{i=0}^{n} [X_i(t) - X_i(0)] + \lambda \sum_{i=0}^{n} \dot{X}_i(t)$$

and the source of randomness disappeared ! This simple remark was exploited by Carlin, Lobo and Viswanathan in [29], and Schied and Schoeneborn in [88] to provide explicit solutions in the deterministic case. It is shown in [29] that there exist unique optimal strategies given by:

$$X_i(t) = ae^{-\frac{n}{n+2}\frac{\gamma}{\lambda}t} + b_i e^{\frac{\gamma}{\lambda}t}$$

where

$$a = \frac{n}{n+2}\frac{\gamma}{\lambda}\left(1 - e^{-\frac{n}{n+2}\frac{\gamma}{\lambda}T}\right)^{-1}\frac{1}{n+1}\sum_{i=0}^{n}[X_i(T) - X_i(0)]$$

$$b_i = \frac{\gamma}{\lambda}\left(e^{\frac{\gamma}{\lambda}T} - 1\right)^{-1}\left(X_i(T) - X_i(0) - \frac{1}{n+1}\sum_{i=0}^{n}[X_i(T) - X_i(0)]\right)$$

The following figures have been chosen to illustrate the impact of the values of the parameters on these solutions. We concentrate on the impact of the number of predators as well as the elasticity or plasticity of the market as given by the relative values of γ and λ governing the relative sizes of the permanent and temporary impacts on the price.

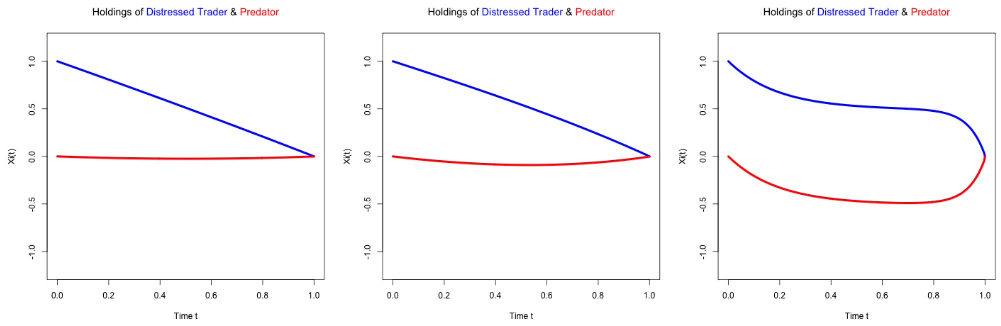

Fig. 1.14. $n = 1$ predator, $\gamma/\lambda = 0.3$ (left), $\gamma = \lambda$ (center), $\gamma = 15.5\lambda$ (right)

1.5.5 Two Period Game Model

For the sake of completeness, we state the conditions and the results obtained in [88] in the case of the extension of the previous model to a two period game. In such an extension, one assumes that:

- The prey has to liquidate $X_0 > 0$ units by time T_1, i.e. $X_0(T_1) = 0$;
- The predators can stay in the game longer $X_i(0) = X_i(T_2) = 0$ for some $T_2 > T_1$ for $i = 1, \cdots, n$;

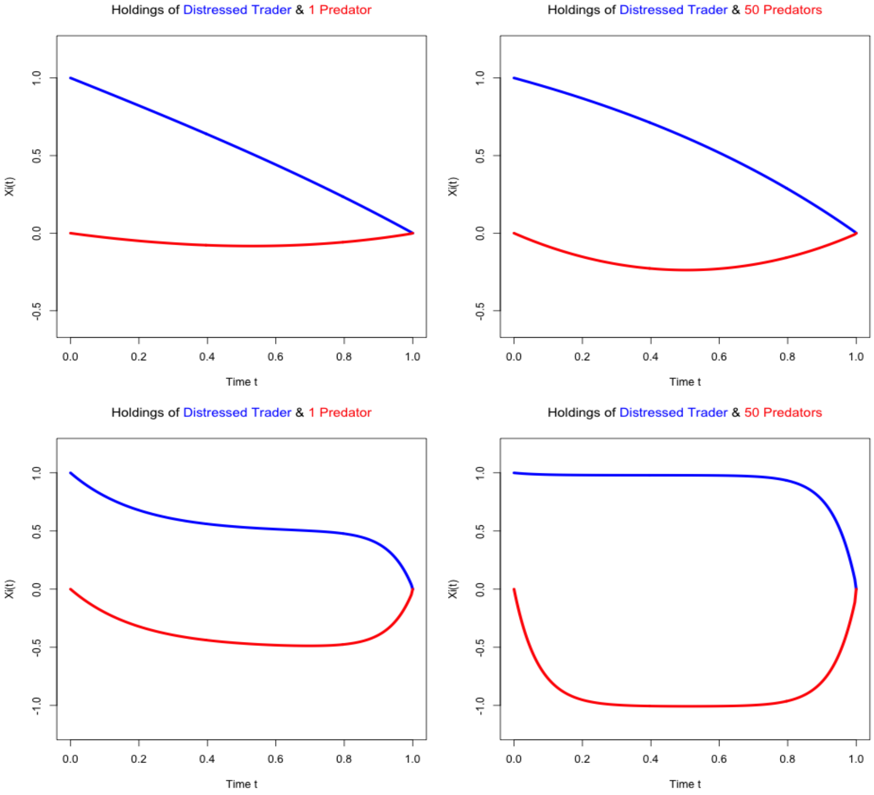

Fig. 1.15. Impact of the Number of Predators: $\gamma = \lambda$ (top), $\gamma = 15.5\lambda$ (bottom)

- The prey does not trade in second period $[T_1, T_2]$, i.e. $X_0(t) = 0$ for $T_1 \leq t \leq T_2$.

Because of the Markovian structure of the problem, the solution is completely determined by the predators' positions at time T_1. The main result is the existence of a unique Nash Equilibrium for deterministic strategies in which ALL the predators have the same position at time T_1

$$X_i(T_1) = \frac{A_2 n^2 + A_1 n + A_0}{B_3 n^3 + B_2 n^2 + B_1 n + B_0} X_0, \qquad i = 1, \cdots, n.$$

This formula is deceptively simple because the coefficients actually depend upon n. However, because they converge as $n \to \infty$ toward constants $\overline{A}_0$, $\overline{A}_1$, $\overline{A}_2$, $\overline{B}_0$, $\overline{B}_1$, and $\overline{B}_2$, the position is written in the form of a rational fraction in n as if the coefficients were constants independent of n. From these formulas, one can derive asymptotic formulas for the expected returns, providing an asymptotic comparison of *stealth* versus *sunshine* trading for some regimes of the ratio γ/λ (which is called GOL for Gamma Over Lambda in the titles of a couple of figures).

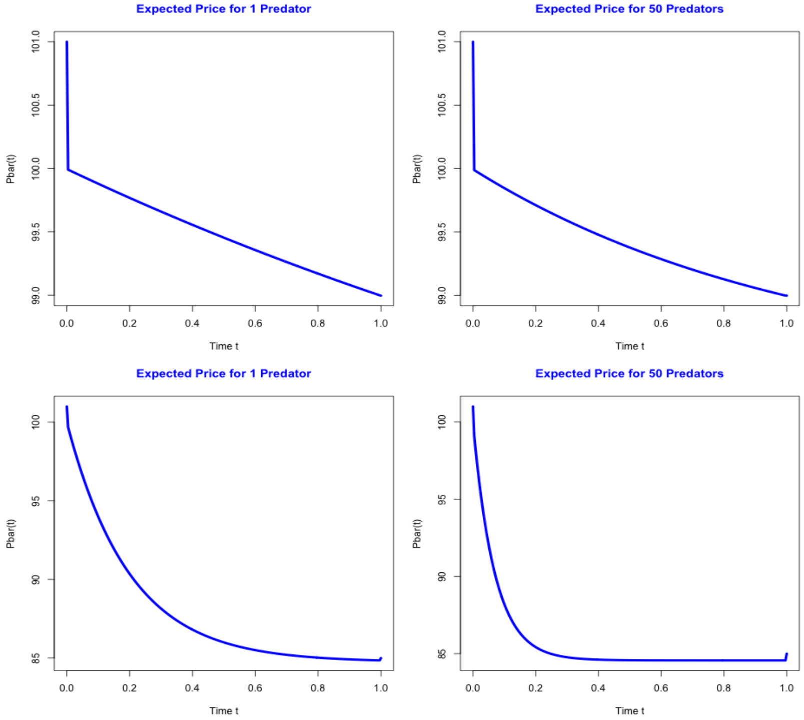

Fig. 1.16. Expected Price: $\gamma = \lambda$ (top), $\gamma = 15\lambda$ (bottom)

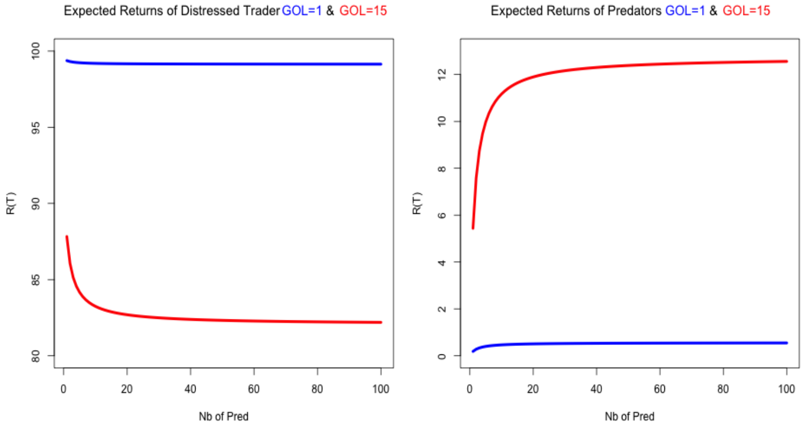

Fig. 1.17. Impact of Nb of Predators on Expected Returns

The deterministic Nash equilibriums identified above can be shown to remain optimal if the players are allowed to use more general adapted trading strategies (analogs of the open loop strategies in the present context). We learned this fact from a private conversation with A. Schied. However, as shown in [33], the situation seems to be different if the admissible strategies are required to be Markovian.

1.5.6 Forensic Analysis

Losses from predatory trading can be significant, and disruptions to the financial system can create havoc. Moreover, the moral and ethical issues associated with this trading practice also contribute to the need for better understanding and detection. This calls for forensic analysis of the known instances of predatory trading, especially the many instances of high frequency trading mishaps discussed in this chapter. However, the application of forensic techniques to the financial markets is still in its infancy.

1.6 Mathematical Models for LOB Dynamics

One of the main goals of LOB modeling is to offer a framework to investigate the impact of orders on execution prices. The framework has to be detailed enough to capture some of the important stylized facts reviewed above. But at the same time, it has to be simple enough to allow for the solution of stochastic optimization problems like: 1) optimal multi-period liquidation strategies against a limit order book; 2) optimal benchmark tracking and slippage control. We discuss some of these difficult optimization problems in the next section. In the meantime, we try to design and analyze detailed but tractable stochastic models for the bid-ask spread and transaction costs which can serve as inputs to the optimization problems.

The following is a short list of works which can be consulted for complements to the presentation of this section. This list is far from exhaustive, but indicative enough of the spirit of these lectures. Equilibrium models can be found in [82, 47, 86, 32], empirical studies in [23, 24, 26, 46, 57], and stochastic dynamic models in [73, 24, 92, 27, 72], while the queuing theory based model discussed in detail below is borrowed from [43].

LOB models cannot be simple, even if one chooses to use only Markovian reduced form models. Indeed, since a LOB is a set of two histograms (a histogram for the bids and a histogram for the asks), the state space has to be large, typically infinite dimensional. We shall denote by $\mathcal{L}$ the state space of order books, and by $L(t)$, a Markov process in $\mathcal{L}$ giving its time evolution.

The most commonly referred to dynamic model for the order book is due to Smith, Farmer, Guillemot and Krishnamurthy, and is known as the SFGK model [92]. In a nutshell, it can be described as a model in which market orders (both buys and sells) arrive according to a Poisson process with a given rate $\mu/2$, cancellations of existing limit orders being modeled in such a way that outstanding limit orders *die* at a rate ν.

In this section, we present a simpler model (still inspired by the SFGK model) for the dynamics of a limit order book. We follow Cont, Stoikov and Talreja [43] in constructing a stochastic model from elementary building blocks. For the purpose of illustration, we depart from the graphical convention used so far to represent order book buy and sell orders with different colors. For mathematical convenience, we posit that the depth available is negative for buy orders and positive for sell orders. Moreover, for the sake of simplicity we assume that the possible price levels are limited and we denote by $\mathcal{P} = \{1, 2, \cdots, n\}$ the price grid in multiples of the price tick. The LOB at time t can be represented in the form of a vector of signed integers, say $L(t) = (L_1(t), L_2(t), \cdots, L_n(t))$ where $|L_p(t)|$ is the volume of active limit orders at price p at time t. According to our convention,

- There are $-L_p(t)$ bid orders at price p if $L_p(t) < 0$;
- There are $L_p(t)$ ask orders at price p if $L_p(t) > 0$.

So the state space of the Markov process used as a model is given by

$$\mathcal{L} = \Big\{ L \in \mathbb{Z}^n;\ L_p < 0 \text{ for } p \leq k,\quad L_p = 0 \text{ for } k < p < \ell,$$
$$\text{and} \quad L_p > 0 \text{ for } \ell \leq p \;\; \text{for some } 1 \leq k \leq \ell \leq n \Big\}.$$

The (best) ask price at time t is then:

$$A_t := (n+1) \wedge \inf\{p;\ 1 \leq p \leq n,\ L_p(t) > 0\}$$

while the best bid price is given by:

$$B_t := 0 \vee \sup\{p;\ 1 \leq p \leq n,\ L_p(t) < 0\}$$

As usual, the mid-price is defined as $p_{mid}(t) = \frac{1}{2}[A_t + B_t]$ and the bid-ask spread by $s(t) = A_t - B_t$

In the next few figures, we describe a few specific transitions in the dynamics of the LOB using the notation introduced above. For the sake of simplicity we only review the most typical events causing the LOB state transitions. The following notation $L^{p\pm v}$ will come handy to denote the transition from the state L of the order book to a state following an order of size (volume) v.

$$L_i^{p\pm v} = \begin{cases} L_i & \text{if } i \neq p \\ L_i \pm v & \text{if } i = p \end{cases}$$

We shall use L^{p-v} only when $L_p \geq v$ in the sequel. For the sake of simplicity, in the following illustrations, we implicitly assume that the requested volume v is smaller than what is available at the relevant level. We already illustrated in Figure 1.3 the impact of a large fill. Figure 1.18 illustrates the impact on the LOB of the arrival of a limit buy order at price level $p < B_t$. The effect is to increase the volume of buy limit orders at price level p

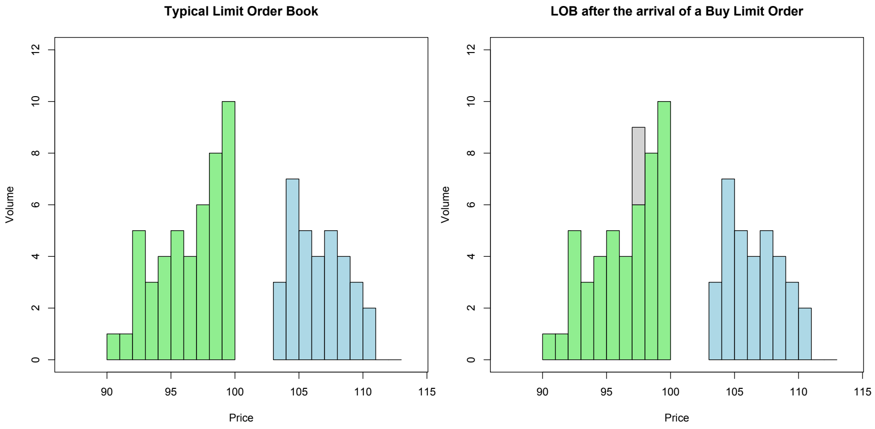

Fig. 1.18. Limit Order Book before (left) and after (right) the arrival of a limit buy order at price level $p < B_t$: $L(t) \hookrightarrow L(t)^{p+v}$

Figure 1.19 illustrates the impact on the LOB of the arrival of a limit buy order at price level $B_t < p < A_t$. The effect is to increase the best bid B_t.

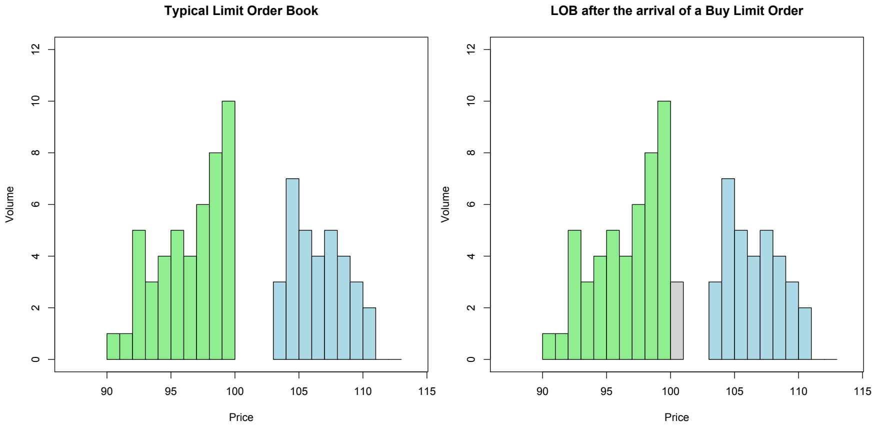

Fig. 1.19. Limit Order Book before (left) and after (right) the arrival of a limit buy order at price level $B_t < p < A_t$: $L(t) \hookrightarrow L(t)^{p+v}$

Similarly, Figure 1.20 illustrates the impact on the LOB of the arrivals of a limit sell order at price level $A_t < p$ (left) and of a limit sell order at price level $B_t < p < A_t$ (right) whose effect is to lower the best ask A_t.

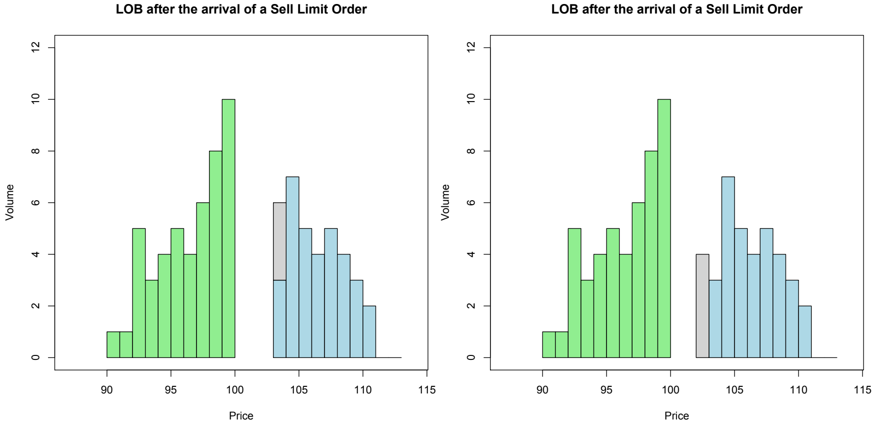

Fig. 1.20. Limit Order Book after the arrival of a limit sell order at price level $A_t < p$ (left) and of a limit sell order at price level $B_t < p < A_t$ (right) whose effect is to lower the best ask A_t

We now switch to the impact of market orders. Figure 1.21 illustrates the transition due to the execution of a market buy order. The effect of such an execution is to decrease the volume at the best ask.

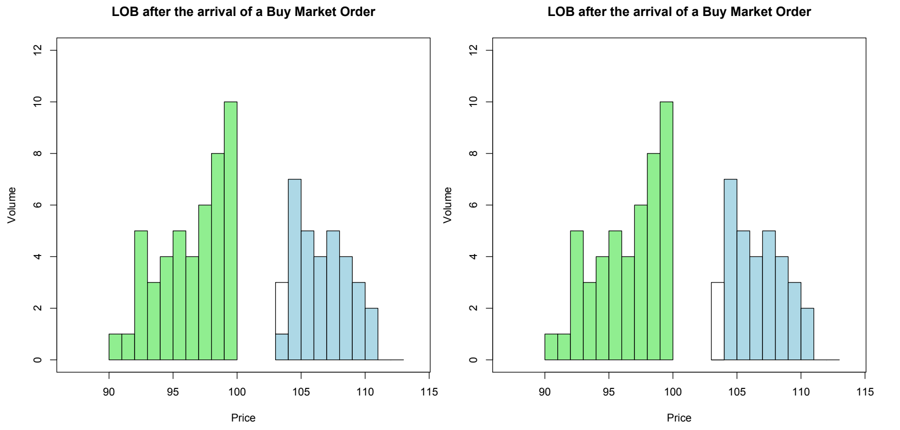

Fig. 1.21. Limit Order Book after the arrival of a market buy order with volume smaller than what was in the book at the best ask (left) $L(t) \hookrightarrow L(t)^{A_t - v}$, and after the arrival of a market buy order with volume equal to what was in the book at the best ask, lowering A_t.

Figure 1.22 illustrates the impacts of cancellations. The cancellation of an outstanding limit buy order at price level $p < B_t$ decreases the quantity at level p: $L(t) \hookrightarrow L(t)^{p-v}$. Similarly, the cancellation of an outstanding limit sell order at price level $p > A_t$ (right) decreases the quantity at level p: $L(t) \hookrightarrow L(t)^{p-v}$.

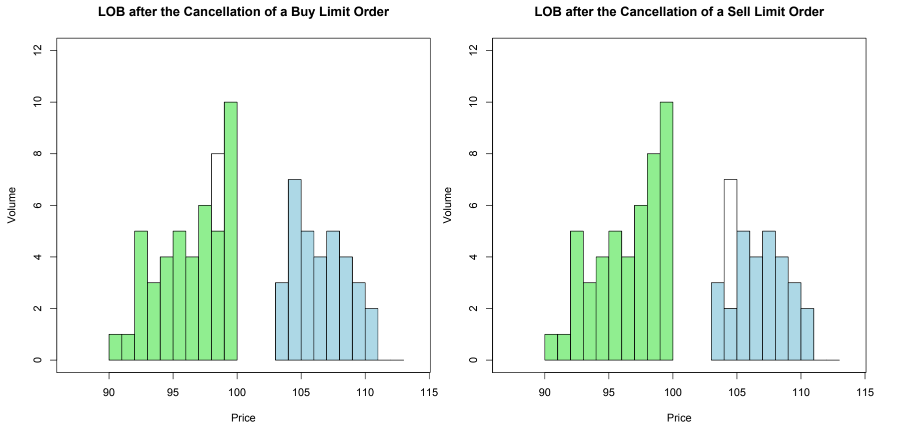

Fig. 1.22. Cancellation of an outstanding limit buy order at price level $p < B_t$ (left) and cancellation of an outstanding limit sell order at price level $p > A_t$ (right).

1.6.1 Probabilistic Assumptions

Following [43] we assume that

- The limit buy and sell orders arrive at a distance of i ticks from the opposite best quote at independent, exponential times with rate $\lambda(i) = Ki^{-\beta}$ for some $K > 0$ and $\beta > 0$;
- The market buy and sell orders arrive at independent, exponential times with constant rate μ;
- Cancellations of limit orders at a distance of i ticks from the opposite best quote occur at a rate proportional to the number of outstanding orders: If the number of outstanding orders at that level is x, then the cancellation rate is $\theta(i)x$;
- The above events are mutually independent.

Under these assumptions, $\underline{L} = [L(t)]_{t\geq 0}$ is a continuous-time Markov chain with state space $\mathcal{L}$ and transition rates:

- $L \hookrightarrow L^{p-1}$ with rate $\lambda(A_t - p)$ for $p < A_t$
- $L \hookrightarrow L^{p-1}$ with rate $\theta(p - B_t)|L_p|$ for $p > B_t$
- $L \hookrightarrow L^{p+1}$ with rate $\lambda(p - B_t)$ for $p > B_t$
- $L \hookrightarrow L^{p+1}$ with rate $\theta(A_t - p)|L_p|$ for $p < A_t$
- $L \hookrightarrow L^{B_t+1}$ with rate μ
- $L \hookrightarrow L^{A_t-1}$ with rate μ

While this is not as obvious as it may seem, this chain does remain in $\mathcal{L}$ if it starts from there, i.e.

$$B_t \leq A_t, \qquad \text{far all } t > 0$$

if it is true at time $t = 0$. Moreover, using ideas from queuing theory and first passage times of birth-and-death processes (usually proven with Laplace transform techniques) one can compute and estimate the probabilities of many events relative to the evolution of the market.

These queuing - like models are rather technical and difficult to manipulate. However, the wide-spread use of diffusion limit approximation results in queuing theory problems gives a glimmer of hope and it may be possible in the near future to incorporate important stylized facts of order flow into these models. The following list is a reasonable set of desirable extensions which would validate some of the empirical results reported in the literature cited earlier:

- Estimation of the limit order arrival rates conditional on distance from e.g. best price on same side;
- Statistics of the active limit orders cancelled and immediately resubmitted;
- Quantification of the aggressiveness of order placement as a function of the depth;
- Frequency of market order arrivals when the spread is large;
- Show that the limit order placement inside the bid-ask spread increases when the depth at best is large;
- (Long-range) autocorrelation of signs of consecutive market orders;

Finally, and this is very unfortunate if not anti-climatic, the dynamic models presented in this section are too cumbersome for the problems of optimal execution which we discuss later.

1.7 Optimization Problems

The goal of a LOB model is to capture enough of the stylized facts of actual real market LOBs to offer a realistic model of the costs of potential transactions and make it possible to develop efficient (if not optimal) trading algorithms. The review of predatory trading of Section 1.5 offers a typical example of the trade-offs needed in order to be able to reach manageable results: the complexity of the trading rules need to be watered down for the costs and benefits of trading to have simple enough dynamics so the optimization problems can be solved. Case in point, we needed to choose the simplest possible form of price impact, as given by the linear form of the Almgren-Chriss' model in order to identify and study equilibriums between preys and predators. And as already mentioned earlier, we shall not discuss here optimal algorithms to split orders into sub-orders to be routed to different exchanges for execution. We refer to [42] for a first attempt in this direction. But again, because of the inherent difficulties in solving the optimization problem, the sophistcation of the order book trading model needs to be reduced to a minimum, in order for the smart routing problem to be tackled with any hope of success.

1.7.1 Generic Optimal Execution Problem

A typical challenge is to schedule the sale of x_0 units of an asset in order to maximize the revenues, using a limited number of market orders. Roughly speaking, the corresponding mathematical problem is to solve the optimization

$$\sup_{\tau_1<\cdots<\tau_n<T} \mathbb{E}\Big[U(\sum_{i=1}^{n} P_B(\tau_i))\Big]$$

where U is a utility function and $\mathbb{E}$ is the expectation over the stochastic model chosen for the dynamics of the LOB L_t. For all practical purposes, the search for optimal strategies and market timing rules is a difficult stochastic control problem in prohibitively high dimension. Its solution is clearly out of reach for the current *technology* and stylized versions based on simplified models are all that can be wished for at this stage of the theory and practice of these models. While the literature on these simplified models is growing at an exponential pace, we review a couple of contributions for the purpose of illustration. We apologize in advance for the bias in our choice of these contributions. While optimal execution is not an exclusive issue of the high frequency markets, the majority of the financial mathematic and financial engineering publications on high frequency trading, still to this day, focus on the optimization of execution or liquidation strategies.

1.7.2 Optimal Execution Tracking a Benchmark

Here we review, without giving any proof, the main results of M. Li's Ph.D. thesis [70].

The goal is to solve an optimal liquidation problem, the originality of the approach being to include *slippage* in the model. The (electronic) broker is penalized for deviating from an agreed upon benchmark, and face a balancing act between liquidating at a best price and avoiding slipping away from the benchmark and having to pay penalties. The results were presented in April 2013 at the UCL Conference *Recent Advances in Algo and HF Trading* where similar, though different results were presented [48]. Note that trading tracking the VWAP (Volume Weighted Average Price) was previously considered in [74]. Since then, a few papers appear with a similar concern about slippage, e.g. [34].

The objective is to sell a quantity $v > 0$ of shares by time $T > 0$, and the trading model involves a mid-price P_t (unaffected price) modeled as a martingale

$$P_t = P_0 + \int_0^t \sigma(u) dW_u, \qquad 0 \le t \le T,$$

the volume $V(t)$ traded in the market up to (and including) time t, the market VWAP $= \frac{1}{V} \int_0^T P_t dV(t)$ where V is the total volume traded on the market during the period $[0, T]$. The fraction of shares still to be executed in the market

$$X(t) = \frac{V - V(t)}{V}$$

will be expressed in the so-called *trade clock*, and after a time change, this fraction of shares becomes

$$X(t) = \frac{T - t}{T} = 1 - \frac{t}{T}$$

which is now deterministic and linear in time: this is a very convenient simplification which we will take advantage of.

Accordingly, we denote by v_t the volume executed by the broker up to time t, and by

$$x_t = \frac{v - v_t}{v}$$

the fraction of shares left to be executed by the broker at time t. We write:

$$x_t = 1 - \ell_t - m_t$$

where we use the notation

- ℓ_t for the (relative) cumulative volume executed through **limit orders**;
- m_t for the (relative) cumulative volume executed through **market orders**.

The broker average liquidation price is given by:

$$\text{vwap} = \frac{1}{v}\int_0^T \left(P_t - \frac{S}{2}\right) dm_t + \left(P_t + \frac{S}{2}\right) d\ell_t$$

and the objective is to minimize the discrepancy between **vwap** and **VWAP**

As explained earlier, in order to have any chance to solve the optimization problem, we need to simplify dramatically the model for the dynamics of the order book. Assuming that the broker controls

- a non-decreasing adapted process $(m_t)_{0\le t\le T}$, and
- a predictable process $(L_t)_{0\le t\le T}$

for the amount of limit orders posted at time t, the part which is not executed instantly being immediately cancelled, so that we have:

$$\ell_t = \int_0^t \int_{[0,1]} y \wedge L_u \; \mu(du, dy) = \sum_{i=1}^{N_t} Y_i \wedge L_{\tau_i}$$

where $\mu(du, dy)$ is a Poisson random measure on $[0, \infty) \times [0, 1]$ whose atoms are the (τ_i, Y_i). Here τ_i should be interpreted as the times at which the crossing market orders arrive, and Y_i should be understood as the size of the market order arriving at time τ_i to cross the existing limit orders. $\mu(du, dy)$ is a Poisson point measure with compensator $\nu_t(du)\nu(t)dt$.

$$x_t = 1 - \int_0^t \int_{[0,1]} y \wedge L_u \; \mu(du, dy) - m_t = 1 - \sum_{i=1}^{N_t} Y_i \wedge L_{\tau_i} - m_t.$$

So the dynamics of x_t are given by

$$dx_t = -\int_{[0,1]} y \wedge L_t \; \mu(dt, dy) - dm_t,$$

with initial condition $x_{0-} = 1$. We now formulate precisely the optimization problem. The goal of the broker is to solve

$$\sup_{(\underline{L},\underline{m})\in\mathcal{A}} \mathbb{E}\Big[U(\text{vwap} - \text{VWAP})\Big]$$

where U denotes the broker's utility function. If we use the CARA (Constant Absolute Risk Aversion) exponential utility, a simple approximation shows that our optimization problem is essentially equivalent to the optimization

$$\inf_{(\underline{L},\underline{m})\in\mathcal{A}} \mathbb{E}\Bigg[\exp\Bigg(-\gamma\left(\frac{S}{2} + \int_0^T [x_u^{L,m} - X(u)]dP_u - S\,dm_u\right)\Bigg)\Bigg],$$

where S denotes the spread (difference between the best ask and the best bid) and $X(u) = (T-u)/T$ represents the fraction of shares left to be executed in the market.

Standard approximation arguments can be used to show that this problem is essentially equivalent to the minimization of a *Mean - Variance* criterion

$$\inf_{(\underline{L},\underline{m})\in\mathcal{A}} \mathbb{E}\bigg[\int_0^T \gamma\frac{\sigma(u)^2}{2}[x_u^{L,m}-X(u)]^2du + S\,m_T\bigg],$$

which look deceivingly simpler because it appears as a linear quadratic stochastic control problem, but which remains extremely challenging because the resulting stochastic control problem is in fact a singular and involves the control of a pure jump process. The value function

$$J(t,x)=\inf_{(\underline{L},\underline{m})\in\mathcal{A}(t,x)} J(t,x,\underline{L},\underline{m})$$

where

$$J(t,x,\underline{L},\underline{m})=\mathbb{E}\bigg[\int_t^T \gamma\frac{\sigma(u)^2}{2}[x_u^{L,m}-X(u)]^2du + Sm_T\bigg].$$

$J(t,x)$ is non-decreasing in t for $x\in[0,1]$ fixed. Indeed, it is easy to see that $\mathcal{A}(t_2,x)\subset\mathcal{A}(t_1,x)$ whenever $t_1\le t_2$.
Unfortunately, as we are about to demonstrate, the optimization problem is **not convex**, because the set $\mathcal{A}$ of admissible controls is not convex.

Indeed, for any number $\ell\in(0,1)$, the two controls $(\underline{L}^1,\underline{m}^1)$ and $(\underline{L}^2,\underline{m}^2)$ by:

$$L_t^1=\mathbf{1}_{\{t\le\tau_1\}}+\sum_{k=2}^{\infty}x_{\tau_{k-1}}\mathbf{1}_{\{\tau_{k-1}<t\le\tau_k\}},\qquad\text{and}\qquad m_t^1=x_{T-}\mathbf{1}_{\{T\le t\}},$$

and:

$$L_t^2=\frac{\ell}{2}\mathbf{1}_{\{t\le\tau_1\}}+\sum_{k=2}^{\infty}x_{\tau_{k-1}}\mathbf{1}_{\{\tau_{k-1}<t\le\tau_k\}},\qquad\text{and}\qquad m_t^2=x_{T-}\mathbf{1}_{\{T\le t\}},$$

are admissible, but the pair $(\underline{L},\underline{m})$ defined by

$$L_t=\frac{1}{2}(L_t^1+L_t^2),\qquad\text{and}\qquad m_t=\frac{1}{2}(m_t^1+m_t^2),$$

is not!

1.7.3 Bibliographical Review of some of the Closest Related Works

Very similar mathematical models have been introduced and studied to determine the optimal amount of reinsurance needed by insurance companies. In these types of insurance models,

- $\mu(dt,dy)$ is a Poisson random measure whose points give the times and sizes Y_t of the claims;

- The insurer pays $Y_t \wedge \alpha_t$ up to a chosen retention level α_t;
- and the remaining excess $(Y_t - \alpha_t)^+$ is passed along to a re-insurer.

The wealth process of the insurance company is given by:

$$X_t = x + \int_0^t p(\alpha_s)ds - \int_0^t y \wedge \alpha_s \, \mu(ds, dy) - \int_0^t dD_s$$

where

- $p(\alpha)$ denotes the insurer net premium (after paying the reinsurance company);
- D_t represents the cumulative dividends paid up to (and including) time t.

The optimization problem is then to compute

$$\sup_{(\alpha_t)_t, (D_t)_t} \mathbb{E}\left[\int_0^\tau e^{-ru} dD_u \right]$$

and the time of bankruptcy is defined as $\tau = \inf\{t \geq 0;\ X_t \leq 0\}$. A similar optimization set-up was used by Jeanblanc and Shyryaev in [61] to study the optimal dividend distribution for a Wiener process. Asmussen, Hojgaard and Taksar generalized this work in [16] to the optimal dividend distribution for diffusion. Finally, Mnif and Sulem used exactly the same set-up to prove existence and uniqueness of a viscosity solution for single contracts in [76], while Goreac treats the case of multiple contracts in [54].

Comparing with the singular stochastic control problem considered in this section we find similarities in that:

- α_t plays the same role as the amount L_t of standing limit orders;
- D_t plays the same role as the cumulative amount m_t market orders.

However, there are also significant differences, among them the fact that

- we work here in with finite horizon, so we need to deal with Partial Differential Equations (PDEs) instead of Ordinary Differential Equations (ODEs);
- we are using a mean - variance criterion;
- finally, we exhibit a classical solution (as opposed to a viscosity solution) as the solution of a hierarchy of ODEs identifying both the value function and an optimal strategy.

1.7.4 Main Result

Under the following technical assumptions (recall that $\nu_t(dy)\nu(t)dt$ is the intensity of the Poisson measure $\mu(dt, dy)$ with $\nu_t([0, 1]) = 1$)

1. $\int_0^T \sigma(t)^2 dt < \infty$
2. $\sup_{0 \leq t \leq T} \nu(t) < \infty$
3. $t \hookrightarrow \frac{\sigma(t)^2}{\nu(t)}(X(t) - x)$ is increasing for each $x \in [0, 1]$

4. $t \hookrightarrow \frac{1}{\nu(t)}\nu_t(\,\cdot\,)$ is decreasing (in the sense of *stochastic dominance*)

It is possible to prove that a form of the dynamic programming principle holds, and that the ensuing Hamilton-Jabobi-Bellman equation takes the form of the following Quasi Variational Inequality (QVI)

$$\min\big[[A\phi](t,x), \partial_t\phi(t,x) + [B\phi](t,x)\big] = 0,$$

where

$$[A\phi](t,x) = S - \partial_x\phi(t,x)$$

and

$$[B\phi](t,x) = \gamma\frac{\sigma(t)^2}{2}[X(t)-x]^2 + \nu(t)\inf_{0\le L\le x}\int_{[0,1]}[\phi(t,x-y\wedge L)-\phi(t,x)]\nu_t(dy)$$

with terminal condition $\phi(T-,x) = Sx$, (notice that $\phi(T,x) = 0$) and boundary condition:

$$\phi(t,0) = \int_t^T \frac{\gamma\sigma(u)^2}{2}X(u)du.$$

In [70] intricate approximation arguments are used to show that a classical solution can be obtained by solving a system of integro-differential equations, and that in the limit:

Theorem 2. *Under the above assumptions, the value function is the unique solution of*

$$\begin{aligned}-\dot J(t,x) = \min\Bigg[&\inf_{0\le y\le x} -\dot J(t,x),\\ &\gamma\frac{\sigma(t)^2}{2}[X(t)-x]^2 + \nu(t)\int_{[0,1]}[J(t,(x-y)\vee \tilde L(t,y)) - J(t,x)]\nu_t(dy)\Bigg]\end{aligned}$$

with

$$J(t,0) = \gamma\int_0^t \frac{\sigma(u)^2}{2}X(u)^2du, \qquad \textit{and} \qquad J(T,x) = Sx,$$

and

$$\tilde L(t,x) = \textit{arg}\min_{0\le y\le x} J(t,y),$$

which is $C^{1,1}$, convex in x for t fixed, non-decreasing in t for x fixed, and satisfying $\partial_x\dot J(t,x)\ge 0$.

Numerical computations based on discrete approximations to the integro-differential system lead to a free boundary (or equivalently a no-trade region) of the form

$$[0,T]\times[0,1] = A\cup B\cup C$$

with

- $A = \{(t,x); \partial_x J(t,x) < 0\} = \{(t,x);\ 0 \le t < \tau_\ell(x)\}$
- $B = \{(t,x); 0 \le \partial_x J(t,x) \le S\} = \{(t,x);\ \tau_\ell(x) \le t \le \tau_m(x)\}$
- $C = \{(t,x); \partial_x J(t,x) = S\} = \{(t,x);\ \tau_m(x) \le t\}$

where

$$\tau_\ell(x) = \inf\{t > 0;\ \partial_x J(t,x) \ge 0\} \quad \text{and} \quad \tau_m(x) = \inf\{t > 0;\ \partial_x J(t,x) \ge S\}$$

because one can prove that

$$\tau_\ell(x) \le T(1-x) \le \tau_m(x).$$

These three regions are represented graphically in Figure 1.23.

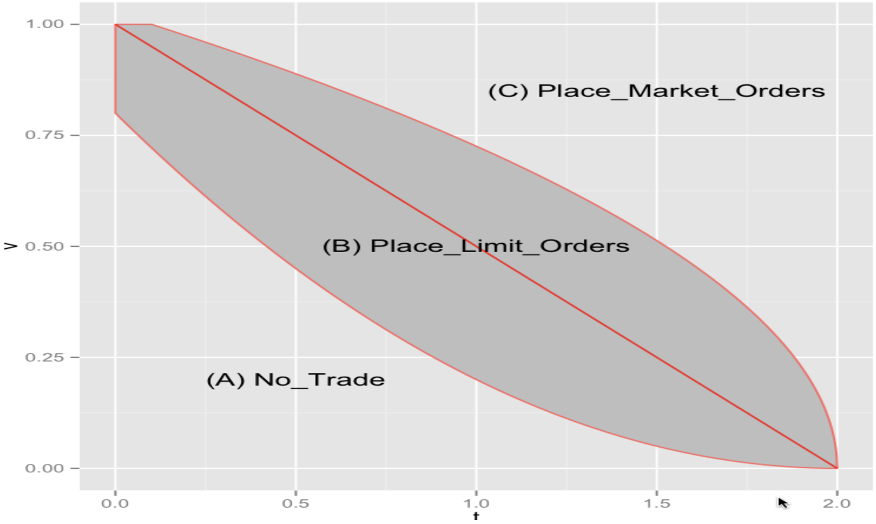

Fig. 1.23. Schematic of the division of the (t,x)-plane into three regions by the (optimal) free boundary.

The corresponding optimal strategy can be summarized as follows:

- If $t > \tau_m(x_t)$ i.e. $(t,x_t) \in C$ (this should never happens if we start from $x = 1$ and use this optimal strategy)
  - **place market orders** $\Delta m_t > 0$ (just enough to get into B)
- If $t = \tau_m(x_t)$ i.e. $(t,x_t) \in \partial C$
  - **place market orders at a rate** $dm_t = -\dot{\tau}_m(x_t)dt$

  (just enough so not to exit B)
- If $\tau_\ell(x_t) \le t < \tau_m(x_t)$ i.e. $(t,x_t) \in B \cup \partial A$
  - **place** $L_t = x_t - \tilde{L}(t)$ **limit orders** (as much as possible without getting too much ahead)
- If $t < \tau_\ell(x_t)$ i.e. $(t,x_t) \in A$
  - **no trade** (but for the same reason as above, this should never happen).

References

1. V.V. Acharyaa and L.H. Pedersen, *Asset pricing with liquidity risk*, Journal of Financial Economics **77** (2005), 375–410.
2. R. K. Aggarwal and G. Wu, *Stock market manipulations*, Journal of Business **79** (2006), 1915–1953.
3. A. Alfonsi and I. J. Acevedo, *Optimal execution and price manipulations in a time-varying limit order book model*, Preprint (2012).
4. A. Alfonsi, A. Fruth, and A. Schied, *Constrained portfolio liquidation in a limit order book model.*, Advances in mathematics of finance, Banach Center Publications, vol. 83, Polish Academy of Sciences, Institute of Mathematics, 2008, p. 925.
5. ———, *Optimal execution strategies in limit order books with general shape functions*, Quantitative Finance **10(2)** (2010), 143157.
6. A. Alfonsi and A. Schied, *Optimal execution strategies in limit order books with general shape functions*, Quantitative Finance (2009).
7. ———, *Optimal trade execution and absence of price manipulations in limit order book models*, SIAM Journal on Financial Mathematics **1** (2010), 490522.
8. ———, *Capacitary measures for completely monotone kernels via singular control*, To appear in SIAM J. Control Optim. (2013).
9. A. Alfonsi, A. Schied, and A. Slynko, *Order book resilience, price manipulation, and the positive portfolio problem*, Tech. report, 2012.
10. R. Almgren, *Optimal execution with nonlinear impact functions and trading-enhanced risk*, Applied Mathematical Finance **10** (2003), 1–18.
11. ———, *Optimal trading with stochastic liquidity and volatility*, SIAM Journal on Financial Mathematics **3** (2012), 163–181.
12. R. Almgren and N. Chriss, *Optimal execution of portfolio transactions*, Journal of Risk, 539.
13. R. Almgren and N. Chriss, *Value under liquidation*, Risk **12** (1999), 61–63.
14. R. Almgren, E. Hauptmann, and H. Li, *Direct estimation of equity market impact*, Risk **18** (2005), no. 7, 58–62.
15. Y. Amihud and H. Mendelson, *Asset pricing and the bid-ask spread*, Journal of Financial Economics **17** (1986), 223–249.
16. A. Asmussen, B. Hojgaard, and M. Taksar, *Optimal risk control and dividend distribution policies. example of excess-of-loss reinsurance for an insurance corporation*, Finance and Stochastics **4** (2000), 299–324.
17. K.H. Bae, H. Jang, and K.S. Park, *Traders choice between limit and market orders: evidence from NYSE stocks*, Journal of Financial Markets **6(4)** (2003), 517538.
18. E. Bayraktar and M. Ludkovski, *Liquidation in limit order books with controlled intensity*, To appear in Mathematical Finance (2011).
19. D. Bertsimas and A. Lo, *Optimal control of execution costs*, Journal of Financial Markets **1** (1998), 150.
20. B. Biais, *Price formation and equilibrium liquidity in fragmented and centralized markets*, Journal of Finance **48** (1993), 157–185.
21. M. Blais and P. Protter, *An analysis of the supply curve for liquidity risk through book data*, International Journal of Theoretical and Applied Finance **13** (2010), no. 6, 821–838.
22. B. Bouchard, N.M. Dang, and C.A. Lehalle, *Optimal control of trading algorithms: a general impulse control approach*, SIAM Journal on Financial Mathematics **2** (2011), no. 1, 404–438.

23. J.P. Bouchaud, *Price impact*, Encyclopedia of Quantitative Finance (Rama Cont, ed.), John Wiley & Sons, Ltd, 2010.
24. J.P. Bouchaud, D. Farmer, and F. Lillo, *How markets slowly digest changes in supply and demand*, Handbook of Financial Markets: Dynamics and Evolution (K. Schenk-Hoppe T. Hens, ed.), North- Holland Publishers, 2008, p. 57160.
25. J.P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, *Fluctuations and response in financial markets: the subtle nature of random price changes*, Tech. report, Science & Finance, Capital Fund Management, 2003.
26. J.P. Bouchaud, M. Mézard, and M. Potters.
27. A. Bovier, J. Cerny, and O. Hryniv, *The opinion game: Stock price evolution from microscopic market modelling*, International Journal of Theoretical and Applied Finance **9** (2006), 91111.
28. M. Brunnermeier and L. Pedersen, *Predatory trading*, Journal of Finance **60** (2005), no. 4, 1825–1863.
29. B. Carlin, M. Lobo, and S. Viswanathan, *Episodic liquidity crises: cooperative and predatory trading*, Journal of Finance **65** (2007), 2235–2274.
30. R. Carmona and K. Webster, *High frequency market making*, Tech. report.
31. ______, *The self-financing equation in high frequency markets*, Tech. report.
32. ______, *Structural relationships in a limit order book*, Tech. report.
33. R. Carmona and J. Yang, *Predatory trading: a game on volatility and liquidity*, Quantitative Finance (2014).
34. A. Cartea and S. Jaimungal, *Optimal execution with limit and market orders*, Tech. report, 2014.
35. U. Çetin, R. Jarrow, and P. Protter, *Liquidity risk and arbitrage pricing theory*, Finance & Stochastics **8(3)** (2004), 311 – 341.
36. CFTC-SEC, *Findings regarding the market events of May 6, 2010*, Report, 2010.
37. A Chakraborti, I.M. Toke, M. Patriarca, and F. Abergel, *Econophysics review: I. empirical facts*, Quantitative Finance **11**.
38. ______, *Econophysics review: Ii. agent-based models*, Quantitative Finance **11**.
39. Constantinides.
40. R. Cont, *Price dynamics in a Markovian limit order market*, Tech. report, 2013.
41. R. Cont and A. de Larrard, *Price dynamics in a Markovian limit order market*, SIAM Journal on Financial Mathematics (2013).
42. R. Cont and A. Kukanov, *Optimal order placement in limit order markets*, Tech. report, 2013.
43. R. Cont, S. Stoikov, and R. Taljera, *A stochastic model for order book dynamics*, Operations Research **58(3)** (2010), 549–563.
44. M. H. A. Davis, V. G. Panas, and T. Zariphopoulou, *European option pricing with transaction costs*, SIAM Journal of Control and Optimization **31** (1993), 470–493.
45. H. Degryse, F. de Jong, M. van Raveswaaij, and G. Wuyts, *Aggressive orders and the resiliency of a limit order market*, Review of Finance **9** (2005), 201–242.
46. D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen, *What really causes large price changes?*, Quantitative Finance **4** (2004), 383397.
47. T. Foucault, O. Kadan, and E. Kandel, *Limit order book as a market for liquidity*, Review of Financial Studies **18(4)** (2005), 11711217.
48. C. Frei and N. Westray, *Optimal execution of a vwap order: A stochastic control approach*, Mathematical Finance (2014), (to appear).
49. A. Fruth, T. Schöneborn, and M. Urusov, *Optimal trade execution and price manipulation in order books with time-varying liquidity*, Tech. report, 2011.

50. M.B. Garman, *Market microstructure*, Journal of Financial Economics **3** (1976), 257–275.
51. J. Gatheral, *No-dynamic-arbitrage and market impact*, Quantitative Finance **10(7)** (2010), 749759.
52. L. Glosten and P. Milgrom, *Bid, ask, and transaction prices in a specialist market with heterogeneously informed agents*, Journal of Financial Economics **14** (1985), 71–100.
53. D.K. Gode and S. Sunder, *Allocative efficinecy of markets with zero intelligence traders: Market as a partial substitute for individual rationality*, Journal of Political Economy **101** (1993).
54. D. Goreac, *Insurance, reinsurance and dividend payment*, Tech. report, 2008.
55. O. Guéant, C.A. Lehalle, and J. Tapia, *Optimal portfolio liquidation with limit orders*, SIAM Journal on Financial Mathematics **3** (2012), 740–764.
56. J. Hasbrouck, *Empirical Market Microstructure: the Institutions, Economics, and Econometrics of Securities Trading*, Oxford University Press, Oxford, 2007.
57. B. Hollifield, R.A. Miller, and P. Sandas, *Empirical analysis of limit order markets*, Review of Economic Studies **71(4)** (2004), 10271063.
58. U. Horst and F. Naujokat, *On derivatives with illiquid underlying and market manipulation*, Quantitative Finance **11** (2011), no. 7, 1051–1066.
59. G. Huberman and W. Stanzl, *Price manipulation and quasi-arbitrage*, Econometrica **72(4)** (2004), 12471275.
60. Securities Industry and Financial Markets Association, *Sifma paper on displayed and non-displayed liquidity*, Tech. report, August 2009.
61. M. Jeanblanc and A.N. Shiryaev, *Optimization of the flow of dividends*, Russian Mathematical Surveys **50** (1995), 257–277.
62. I. Kharroubi and H. Pham, *Optimal portfolio liquidation with execution cost and risk*, SIAM Journal on Financial Mathematics **1** (2010), 897–931.
63. A. Kirilenko, A. Kyle, M. Samadi, and T. Tuzun, *The flash crash: The impact of high frequency trading on an electronic market*, Tech. report, University of Maryland, 2010.
64. P. Kratz and T. Schöneborn, *Optimal liquidation in dark pools*, Tech. report, 2010.
65. A. Kyle, *Continuous auctions and insider trading*, Econometrica **53** (1985), no. 6, 1315–1336.
66. A. Kyle and S. Viswanathan, *How to define illegal price manipulation?*, American Economic Review: Papers & Proceedings **98** (2008), 274–279.
67. S. Laruelle, C.A. Lehalle, and G. Pagès, *Optimal split of orders across liquidity pools: a stochastic algorithm approach*, SIAM Journal on Financial Mathematics **2(1)** (2011), 1042–1076.
68. C-A. Lehalle and S. Laruelle, *Market microstructure in practice*, World Scientific, 2014.
69. C.A. Lehalle, *Market microstructure knowledge needed to control an intra-day trading process*, Handbook on Systemic Risk (J.P. Fouque and J. Langsam, eds.), Cambridge University Press, 2013.
70. M. Li, *Optimal execution tracking a benchmark*, Ph.D. thesis, Princeton University, Department of Operations Research and Financial Enginnerring, 2012.
71. R. Liu and J. Muhle-Karbe, *Portfolio choice with stochastic investment opportunities: a user's guide*, Tech. report, Proc. 1st Princeton Summer School in Mathematical Finance, 2013.
72. H. Luckock, *A steady-state model of the continuous double auction*, Quantitative Finance **3** (2003), 385404.
73. S. Maslov and M. Mills, *Price fluctuations from the order book perspectiveempirical facts and a simple model*, Physica A **299** (2001), 234246.

74. J. McCulloch and V. Kazakov, *Optimal VWAP trading strategy and relative volume*, Quantitative Finance Research Centre (2007), 31.
75. H. Mittal, *Are you playing in a toxic dark pool? A guide to preventing information leakage*, Journal of Trading **3** (2008), no. 3, 20–33.
76. M. Mnif and A. Sulem, *Optimal risk control and dividend policies under excess of loss reinsurance*, Tech. report, 2005.
77. R. Naes and J.A. Skjeltorp, *Order book characteristics and the volume-volatility relation: Empirical evidence from a limit order market*, Journal of Financial Markets **9(4)** (2006), 408–432.
78. P. Oasis, *High frequency markets analysis from order flows and interactions*, 2012.
79. A. Obizhaeva and J. Wang, *Optimal trading strategy and supply/demand dynamics*, Journal of Financial Markets (2005).
80. M. O'Hara, *Market microstructure theory*, Basil Blackwell, 1995.
81. ______, *What is a quote?*, The Journal of Trading **52** (2010), 10–16.
82. C. Parlour, *Price dynamics in limit order markets*, Review of Financial Studies **11(4)** (1998), 789816.
83. C.A. Parlour and D.J. Seppi, *Limit order markets: a survey*, Handbook of Financial Intermediation & Banking (A.W.A. Boot and A.V. Thakor, eds.), 2008.
84. M. Potters and J.P. Bouchaud, *More statistical properties of order books and price impact*, Physica A: Statistical Mechanics and its Applications **324(1-2)** (2003), 133140.
85. Predoiu, Shaikhet, and Shreve, *Optimal execution in a general one-sided limit order book*, SIAM Journal on Financial Mathematics **2** (2012), 183212.
86. I. Rosu, *A dynamic model of the limit order book*, Review of Financial Studies **22** (2009), 46014641.
87. A. Schied, *Robust strategies for optimal order execution in the Almgren–Chriss framework*, Applied Mathematical Finance (2011).
88. A. Schied and T. Schöneborn, *Risk aversion and the dynamics of optimal trading strategies in illiquid markets*, Finance and Stochastics **13** (2009), 181–204.
89. A. Schied, T. Schöneborn, and M. Tehranchi, *Optimal basket liquidation for CARA investors is deterministic*, Applied Mathematical Finance **17** (2010), 471–489.
90. T. Schöneborn and A. Schied, *Liquidation in the face of adversity: stealth vs. sunshine trading*, Tech. report, 2009.
91. S. Shreve and M. Soner, *Optimal investment and consumption with transaction costs*, The Annals of Applied Probability **4** (1994), 609–692.
92. E. Smith, J.D. Farmer, L. Gillemot, and S. Krishnamurthy, *Statistical theory of the continuous double auction*, Quantitative Finance **3** (2003), 481514.
93. P. Weber and B. Rosenow, *Order book approach to price impact*, Quantitative Finance **5** (2005), 357–364.
94. I. Zovko and D. Farmer, *The power of patience; a behavioral regularity in limit order placement*, Quantitative Finance **2** (2002), 387392.