The Limit Order Book: A Survey

Martin D. Gould,1,* Mason A. Porter,1 Stacy Williams,2
Mark McDonald,2 Daniel J. Fenn,2,3,4 and Sam D. Howison1

1Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX1 3LB, UK
2FX Research and Trading Group, HSBC Bank, 8 Canada Square, London E14 5HQ, UK
3Mathematical and Computational Finance Group, Mathematical Institute, University of Oxford, Oxford OX1 3LB, UK
4CABDyN Complexity Centre, University of Oxford, Oxford OX1 1HP, UK

As the pricing mechanism in more than half the world’s financial markets, the limit order book has recently been the focus of a great deal of published literature in a wide range of disciplines. In this survey, we present a mathematical description of the price matching algorithm at the heart of limit order trading, and highlight some of the key publications – both empirical and theoretical – which have advanced understanding of the process to date. By examining existing models of limit order markets, we identify some of the key unresolved questions and difficulties facing researchers of limit order trading today.

I. INTRODUCTION

Determining the price at which to conduct trade is an age-old problem. The first (albeit primitive) pricing mechanism dates back to the Neolithic era, when people met in physical proximity in order to agree upon mutually beneficial exchanges of goods and services [10]. Subsequently, as markets have evolved through the ages, increasingly complex mechanisms have played a role in determining prices. In the highly competitive and relentlessly fast-paced markets of today’s financial world, it is the limit order book that matches buyers and sellers to trade at an agreed price in more than half of the world’s financial markets [35]. A recent study [27] has claimed that a limit order market can generate almost 65% more trading activity than if the same market participants were to trade in a Walrasian market,1 and thanks to technological advances, traders all over the world now have real-time access to the current limit order book, providing buyers and sellers alike “the ultimate microscopic level of description” [9].

Surprisingly little is currently known about the complex systems underpinning the mechanics of limit order trading, especially considering how widely the mechanism is used in the financial community. There are obvious practical advantages to better understanding the dynamics of the limit order book, such as gaining a clearer insight into the relative financial merits of limit versus market orders in given situations (which has itself been explicitly explored in [21]). In this survey we review some of the key mathematical ideas that have emerged from limit order trading in recent years and explore the extent to which a selection of theoretical economic attributes can be derived mathematically – or, at the very least, observed empirically – from the mechanism. We also consider the extent to which, by its very existence, limit order book trading is affecting markets. We conclude the survey by examining some recent attempts to produce a mathematical model of limit order trading and discuss the strengths and limitations of the models as they stand.

The remainder of the survey is organised as follows: In Section II, we discuss some qualitative aspects of limit order trading and introduce some of the formal definitions related to the process. These definitions are the building blocks for a rigorous mathematical description of limit order trading, which we formulate in Section III. In Section IV, we discuss some of the practical aspects of limit order trading and examine the mathematical difficulties that arise from attempting to quantify them. In Section V, we examine how empirical studies have played an important role in deepening understanding of the limit order book, highlighting both consensus and disagreement across different publications. In Section VI, we discuss how some specific trading strategies exploit the unique features of limit order trading. A selection of models which have been proposed in the published literature are examined in Section VII. Section VIII contains our conclusions and discussions of the key unresolved issues in the field.

*Corresponding author: gouldm@maths.ox.ac.uk

1 A Walrasian market is one in which all market participants send their desired buy or sell orders to a specialist, who then determines the market value of the asset being traded by selecting the price that will maximise the volume of trade.
II. LIMIT ORDER TRADING

Before limit order trading grew in popularity, almost all financial trades took place in quote driven marketplaces. In such a setup, a handful of large market makers centralise “buy” and “sell” orders by publishing the price at which they would be willing to buy and price at which they would be willing to sell the asset being traded. Naturally, the “sell” price will always be higher than the “buy” price. This reflects the “fee” that the market maker charges for the service they offer – i.e., providing guaranteed liquidity to the market. Any other market participants who want to buy or sell the asset being traded only have access to the prices made publicly available by these market makers, and the only action available to a market participant who wishes to buy (resp. sell) the asset being traded is to submit an order to immediately buy (sell) at the lowest (highest) available price in the market.

A limit order driven market is, on the other hand, much more flexible because every market participant has the option of posting two types of buy (sell) orders: “Impatient” market participants can choose to immediately buy (sell) a specified quantity of the asset being traded at the lowest (highest) available price in the market, in much the same way as they could in the quote driven system. However, “patient” market participants also have another option: they can post an order to the limit order book, stating an intention to buy (sell) a specified quantity of the asset being traded at a specified price. This order will then remain in the limit order book either until it is cancelled by its owner or until the time that it is the best offer among all those in the limit order book and is matched to an incoming order from an “impatient” market participant. In this way, an “impatient” buy (sell) order contains only one piece of information: the quantity to be purchased (sold). A “patient” order, however, contains two pieces of information: the desired quantity to be purchased (sold) and the price at which to do so. At any given time the limit order book is a current list of all those “patient” market participants who have declared a desire to buy (sell) the asset being traded along with the price at which they are willing to do so.

In almost all of the published literature on limit order books, a universal terminology has been adopted: the orders that we have above called “impatient” orders are known as “market” orders, and the orders that we have above called “patient” orders are known as “limit” orders. However, there are several problems with defining the terminology in this way. The first is that some limit order books require that even “impatient” orders specify an execution price in order to safeguard their owners from sudden swings in market price causing “impatient” market participants to pay far worse prices than they had initially intended. We discuss this issue in more detail in Section IV. Deeming an order to be a market order if and only if it specifies just a trade quantity (and not a price) is, therefore, inaccurate. Furthermore, it is possible to submit an order that contains both a trade quantity and a price in such a way that it is matched immediately. For example, submitting a buy (sell) order for a quantity and price that is already offered for sale (purchase) in the limit order book will result in an immediate match, so deeming any order to be a limit order if and only if it specifies both a trade quantity and a price is also inaccurate.

To overcome this ambiguity, Bouchaud et al. [8] suggested that orders should be partitioned according to whether or not they result in an immediate match, irrespective of what their owner intended them to do. They call any order that is matched immediately an “effective market order” and any order that is not matched immediately an “effective limit order.” This is the convention that we will adopt throughout the survey, but to avoid endless repetition of the word “effective,” we will use the phrase “limit order” to mean “effective limit order,” and the phrase “market order” to mean “effective market order,” unless we explicitly state otherwise. we now formalise these definitions:

Definition. A market order of size \( \omega \) is an order to buy (sell) \( \omega \) units of the asset being traded at the lowest (highest) available price in the market.

Definition. A limit order of size \( \omega \) at price \( p \) is an order to buy (sell) \( \omega \) units of the asset being traded at the specified price \( p \).

To recap, a market order submitted at time \( T \) is always matched at time \( T \), whereas a limit order submitted at time \( T \) is never matched at time \( T \) – instead, it is posted to an electronic trading system where it is stored

---

2 The concept of liquidity is somewhat difficult to define formally – Kyle [25] instead identifies the three key properties of a liquid market to be tightness (“the cost of turning around a position over a short period of time”), depth (“the size of an order flow innovation required to change prices a given amount”) and resiliency (“the speed with which prices recover from a random, uninformative shock”).

3 Notice also that it is possible for a “patient” buy (sell) market order to be submitted at a price strictly higher (lower) than the best price offered at that time in the limit order book. In this case, we will assume that the trade-matching algorithm steps in and changes the price of such an order to be equal to the best price available in the limit order book before performing the matching. This is then in line with considering such an order to be a “traditional” market order (i.e., matched at the best price).
until either matched to an incoming market order at some \( t > T \) or cancelled (again at some \( t > T \)). Cancellations generally occur because the seller no longer wishes to trade at the stated price, but rules governing the marketplace can also lead to the cancellation of limit orders. For example, most equity markets forcibly cancel all existing limit orders at the end of the trading day. In the above setup, if a trader submits an order that is only partially matched immediately, we consider the order to consist of two parts: a market order part (which is matched immediately) and a limit order part (which is posted to the limit order book).

We now give a precise definition for an active limit order:

**Definition.** An active limit order at time \( t \) is a limit order that has been submitted at some time \( T < t \), but which has not been fully filled or cancelled by time \( t \).

It is precisely the active limit orders in a market that make up the limit order book:

**Definition.** The limit order book \( L(t) \) is the collection of all active limit orders in the market at time \( t \).

It is important to note that not all major financial markets have switched to a limit order driven system. While there are relatively few markets that still operate under a purely quote driven mechanism, there are a significant number that operate some bespoke hybrid system. These include the New York Stock Exchange, NASDAQ, and the London Stock Exchange [14]. On the other hand, many major exchanges (such as Euronext; the Helsinki, Hong Kong, Swiss, Tokyo, and Toronto Stock Exchanges; and the Australian Securities Exchange) now operate as pure limit order markets [26, 35].

### III. MATHEMATICAL PRELIMINARIES

The following are some of the key definitions and formalities that will be used throughout the remainder of the survey.

**Definition.** The depth profile at price \( p \) and at time \( t \), denoted \( n(p,t) \), is the density of the total volume of the asset being traded that is offered via limit orders at price \( p \) and at time \( t \).

**Definition.** The quantity available at price \( p \) and at time \( t \), denoted \( N(p,t) \), is the amount available in the limit order book \( L(t) \) at price \( p \). Hence, \( N(p,t) = n(p,t)dp \).

We adopt the convention that the depth profile (and quantity available) is negative for buy orders and positive for sell orders.

**Definition.** The minimum order size, denoted \( \sigma \), is the smallest quantity of the asset that can be traded. All orders must be for sizes that are integer multiples of \( \sigma \) - i.e., \( \omega \in \{k\sigma | k \in \mathbb{Z}\} \) (where \( k \) is negative for buy orders).

**Definition.** The bid price at time \( t \), denoted \( b(t) \), is equal to the highest stated price among “buy” limit orders in the limit order book \( L(t) \).

**Definition.** The ask price at time \( t \), denoted \( a(t) \), is equal to the lowest stated price among “sell” limit orders in the limit order book \( L(t) \).

In a limit order market, \( b(t) \) is precisely the highest price at which it is possible to sell a quantity of at least \( \sigma \) of the asset being traded, and \( a(t) \) is precisely the lowest price at which it is possible to buy a quantity of at least \( \sigma \) of the asset being traded, at time \( t \).

**Definition.** The bid-ask spread at time \( t \), denoted \( s(t) \), is the difference between the ask price at time \( t \) and the bid price at time \( t \). More succinctly, \( s(t) = a(t) - b(t) \).

**Definition.** The mid price at time \( t \), denoted \( m(t) \), is the arithmetic mean of the ask price at time \( t \) and the bid price at time \( t \). More succinctly, \( m(t) = \frac{a(t) + b(t)}{2} \).

**Definition.** The tick size \( dp \) is the smallest price interval between different orders that is permissible in the market. Note that \( dp \) is not infinitesimally in general.\(^5\)

---

\(^4\) This varies greatly from market to market. Expensive shares are sometimes traded in single units, yet in many foreign exchange markets, the currency pair \( XXX/YYY \) is traded in “millions of XXX.”

\(^5\) For example, if the best bid is currently \$1.2672 and \( dp \) is \$0.0001 in the market, then the smallest permissible limit order that would beat the best bid is \$1.2673 (and indeed any limit order would have to be submitted at a price with exactly four decimal places). In equity markets, \( dp \) generally scales up with the current price of the underlying asset. Common examples of \( dp \) in foreign exchange markets are 0.0001p for \( GBP/USD \) trades, but 0.01 for \( USD/JPY \) trades. \( dp \) is typically constitutes a small number of basis points (where a basis point is 0.0001 units) in most foreign exchange markets.
The minimum order size \( \sigma \) and the tick size \( dp \) are commonly called the limit order book’s resolution parameters. In order to attract traders, many new trading platforms offer smaller values of \( \sigma \) and \( dp \). For example, it is often possible to post limit orders in amounts of one hundredth of a unit in new foreign exchange trading platforms (e.g., a USD/XXX order size could be specified up to the nearest cent). In these exchanges, however, individual traders have the option to specify a minimum match size – that is, the minimum amount of currency that a limit order has to offer before it is considered for matching to their order. This prevents large market orders being matched to hundreds of tiny limit orders, which would cause a large amount of administrative work to process the trades. As we discuss in Section VI, however, changing resolution parameters can have dramatic (and often unexpected) consequences for the limit order book.

Figure 1 shows a schematic representation of a limit order driven market at some instant in time, highlighting the roles that the above definitions play.

If a new market order arrives at time \( t \), it is matched to the best active limit order (of opposite type) in the limit order book \( L(t) \). However, if for example a new sell market order is placed for 10 shares, and immediately prior to this arrival 7 shares are available at \( b(t) = 1.0312 \), \( dp = 0.0001 \), and 15 shares are available at \( 1.0311 \), then the market participant who submitted the market order actually only sells 7 shares at \( 1.0312 \). They are forced to sell the remaining 3 shares at the worse price of \( 1.0311 \). Thus, the new bid price immediately after a sell market order of size \( x \) has been submitted is the smallest \( p' \) that solves the equation

\[
x \geq \sum_{p=p'} b(t) n(p,t) dp.
\]  

(1)

The equivalent equation for finding the new ask price after a new buy market order of size \( x \) is

\[
x \geq \sum_{p=a(t)}^{p'} n(p,t) dp.
\]  

(2)
IV. LIMIT ORDER MARKETS

Incomplete Sampling

An important consideration for any study on limit order markets is the extent to which the state of the limit order book, as reported by the electronic trading platform, really reflects people’s trading intentions. Bouchaud et al. [8] highlighted the fact that a typical snapshot of the limit order book at a given time is very sparse, containing few orders that are very far from the current bid/ask. This should not be seen as an indication that few people wish to trade at these prices, however – it is merely an indication that they have not announced any intention to trade at them. Although it is impossible to know specific trading strategies exactly, a likely explanation for this phenomenon is that traders see little point in placing limit orders at extreme values, viewing the probability of them being matched to be too low. As discussed in Section V, this behaviour could explain the existence of the empirically observed “hump” in the average state of the limit order book.

Some market participants choose not to submit limit orders at all. These traders instead watch how the values of $b(t)$ and $a(t)$ evolve with time and place market orders when certain criteria are met. They know in advance what these criteria are, of course, so in some respects their behaviour is similar to placing “imaginary” limit orders, but without announcing this to the market in any way until $a(t)$ or $b(t)$ are at such a level that their intentions become solidified in a real order.

Both of the above examples show that at any given moment $L(t)$ is actually just a subset of people’s trading intentions – that is, $L(t)$ reflects the subset of trading intentions that market participants have announced up to time $t$. It is, however, impossible to quantify trading intention that has not been announced via the limit order book in some quantitative way, so we restrict our attention to the disclosed limit order book $L(t)$.

Rival Schools

In traditional economics literature, market behaviour is usually linked to abstract concepts such as “rational” agents attempting to maximise their “utility” by making trades in markets driven by “information.” While these ideas certainly give some insight into the human nature of trading, they have recently come under fire. For example, Gode and Sunder [20] highlighted how utility maximisation is often inconsistent with direct observations of individual behavior, and Smith et al. [37] noted that real traders do not all instantaneously update their limit orders with the arrival of every new piece of information (which they would need to do under perfect rationality). The difficulty with using largely qualitative concepts such as “perfect rationality” in quantitative research has recently led to the emergence of the zero intelligence approach.

In their seminal paper [20], Gode and Sunder claimed that in a limit order driven market populated exclusively by zero intelligence agents, there is still convergence of the price process to the theoretical equilibrium. Since then, numerous other papers (e.g., [28, 37]) have demonstrated that agent-based models built upon some form of zero intelligence setup can exhibit statistical properties very similar to those found in real market data. The strength of such results has come under question by other authors (e.g., [13]) who believe that some elementary learning on behalf of the agents is required for price convergence to occur in real markets, but nonetheless it has been shown that the assumptions associated with perfect rationality are not strictly necessary in quantitative finance research. Much like perfect rationality, zero intelligence is an extreme simplification, but zero intelligence has the appeal of being a more easily quantifiable concept than perfect rationality. In essence, the strength of taking a zero intelligence approach is the ability to model the overall effect of many agents acting independently, without having to predict how the individual agents’ behaviours contribute to the aggregate observation. The approach can, of course, be extended by introducing some simple boundedly rational behaviours, but rather than attempting to quantify exactly how rationality or information directly affects a price, it is sufficient to note that they can affect order flow rates, which are precisely the “forces” driving the evolution of the price process.

---

6 Arbitragers are a key example of this. Their strategies depend on simultaneously buying and selling assets, in an attempt to make instant profit. Limit orders are of little use to them, due to the uncertainty of when the limit order will be matched (if ever).
7 i.e., agents who place market and limit orders that obey the structure of the market, but do so entirely at random.
8 We explore in Section VII how some authors have attempted to quantify perfect rationality for modelling purposes and discuss the often highly unrealistic assumptions that such formulations require. A more detailed history of such modelling can be found in [18].
9 A piece of information might, for example, make traders more impatient (i.e., more demanding of immediate liquidity, as discussed in Section V), which in turn increases the flow of market orders.
Coupling between $a(t)$ and $b(t)$

Even when agents act with zero intelligence, $a(t)$ determines the boundary condition for effective buy limit order placement because any buy limit order placed at or above $a(t)$ is simply converted to a buy market order by the trade matching algorithm. $b(t)$ plays a similar role for sell limit orders. Furthermore, there are multiple interacting processes which affect $a(t)$ and $b(t)$:

- Buy market orders and sell limit orders directly affect $a(t)$
- Sell market orders and buy limit orders directly affect $b(t)$
- Cancellations of sell limit orders affects $a(t)$ and cancellations of buy limit orders affects $b(t)$

There is, therefore, a strong coupling between $a(t)$ and $b(t)$. Smith et al. [37] stated that it is precisely this non-linear coupling between $a(t)$ and $b(t)$ which makes the study of the limit order book a rich and difficult mathematical problem.

Priority

As is clear from Figure 1, it is possible to have limit orders from different market participants at the same price at any given time in the limit order book. Much like priority is given to those limit orders with the best (i.e., highest bid or lowest offer) price when a market order is placed, there is also a priority system for orders within each individual price.

By far the most common priority mechanism currently used in financial markets is price-time. That is, priority is first given to those limit orders with the best price, and ties are broken by selecting the limit order that was placed first among those limit orders with the best price. We explore in Section VI how a better understanding of the price-time mechanism has led to the evolution of trading strategies in the limit order book.

An alternative priority mechanism, however, is price-size, in which priority is first given to those limit orders with the best price and ties are broken by selecting the limit order of the largest size among those limit orders with the best price. NASDAQ OMX Group is currently conducting research into price-size priority, with a view to roll out a new equity trading platform using the mechanism in mid-2010 [1].

Both systems clearly have advantages: price-time priority encourages traders to act quickly in order to get their quotes at the front of the queue, and price-size priority rewards traders for placing large limit orders and thus for providing greater liquidity to the market. It will be interesting to see whether NASDAQ’s new trading platform paves the way for a whole host of new price-size prioritised markets in a landscape that is currently dominated by price-time priority systems.

Hidden Liquidity

In recent years, some electronic limit order books have permitted the use of what are colloquially known as “iceberg orders.” An iceberg order is a type of limit order that specifies not only a total size and price but also a visible size. All other market participants then only see the visible size. How the “hidden” quantity is dealt with varies greatly from exchange to exchange. In some cases, once a quantity of at least the visible size is matched to an incoming market order, another quantity equal to the visible size becomes visible, with a time priority position equal to a standard limit order being placed at this time. This sort of iceberg order is broadly equivalent to a trader watching the market very carefully and submitting a new limit order at the same price and size at the exact moment that their previous limit order is matched to an incoming market order. The only difference occurs when the trader is the only market participant making a limit order at that price – by submitting an iceberg order, if a large market order arrives then the trader will match his visible limit size and then a portion (perhaps all) of his hidden limit amount to the market order. In contrast, a trader submitting small but entirely visible duplicate limit orders would have a large incoming market order matched only to the amount stored in his limit order at the time, with the rest of the incoming market order instead being matched to the next best limit order available.

Other exchanges have an alternative structure for iceberg orders. Whenever a quantity equal to at least the visible amount of an iceberg order is matched to an incoming market order, the rest of the order (i.e., the portion of the hidden component that is not also matched to the same incoming market order) is cancelled. In this way, if a trader is the only market participant who is offering a limit order at a given price, they can match incoming market orders of a bigger size than is initially apparent in the market (because the market only displays the visible portion of the limit order), but otherwise the iceberg order behaves like any other
order. This is the system currently used by the Reuters trading platform (where the hidden quantity is known as the “more quantity”) [2].

Other trading platforms (e.g., Hotspot in the United States) allow traders to place entirely hidden limit orders. These orders are given priority behind any visible or iceberg orders at their price, but they give traders the ability to make limit orders at a specified price without revealing any information to the market.

As the above examples illustrate, many exchanges allow market participants to hide the extent of their intentions to trade at the cost of paying some penalty in terms of priority. Consequently, iceberg orders pose a significant problem in analysing market data. Inferences must be drawn from what is observed in the market, but by allowing iceberg orders the trading platforms permit significant (and complex) structure to exist in an unobservable manner. As numerous authors comment (eg. [6, 8, 23]), the presence of hidden orders in a market makes the study of the limit order book all the more difficult. Occasionally, market data allows inference about when hidden orders have affected the market, but this is not possible to do in general.

There is also an entirely different sense in which liquidity can be “hidden” in the market. In order to speed up the trading process, financial institutions allocate credit to other institutions with whom they are happy to trade, but a large bank might feel that a small, unknown individual—or worse, an individual with a bad track record—is too risky to have any credit assigned. This information is passed on to the exchange, and any limit order placed by either of these parties will not appear visible in the limit order book viewed by the other party. Any market orders made by either party bypass any limit orders made by the other. Although such trades account for only a small fraction of activity in the market, they highlight the fact that it is possible for market orders to be matched to limit orders that are not at the current bid/ask price. This again provides further difficulty in formulating an accurate mathematical description of matching in the limit order book. To our knowledge, no tractable models that are capable of dealing with either of the “hidden liquidity” scenarios discussed exist in the literature.

Some Specific Limit Order Markets

In practice, most real financial markets are based on minor variations of the general principles described above. We now briefly discuss the setups of some specific markets.

- **The London Stock Exchange** state that their flagship order book SETS “executes millions of trades a day at millisecond latencies” [3]. As well as allowing hidden orders (as discussed above), SETS also has a slightly unusual pricing mechanism. For the majority of the trading day, the standard limit order book mechanism is used, but for the first and last fifteen minutes of the trading day a “price finding” auction takes place. During this time, all market participants can view and place limit orders as usual, but market orders that are placed during the “price finding” period are not executed immediately. Instead, they are stored and displayed to all traders until the “price finding” period is over, at which point they are all simultaneously matched to limit orders in the limit order book as it stands at that time (with a time priority mechanism to ensure that market orders that were placed first get the best prices). The purpose of this seemingly odd practice is to allow all market participants to observe the “discovery” of the price without any trades actually happening until this process is complete.

- In both of the major electronic trading platforms used for foreign exchange trades in London (i.e., Reuters and EBS), all orders must come with a specified limit price. This safeguards market participants against sudden swings in the market, which (due to latency) can occur in the short period of time between an order being submitted by a trader and that order reaching the exchange. Incoming orders are assessed when they reach the exchange. If they can be wholly (partially) matched at the stated limit order price or better, they are immediately wholly (partially) converted into a market order. Any quantity that cannot be matched immediately is converted into a limit order at that size and price and is placed in the limit order book.

- Traders sometimes submit market orders for quantities larger than the (visible) amount available among all limit orders at that time.11 For this reason, traders submitting market orders are requested to

---

10 The term “latency” is used to describe the delay experienced by market participants between the time that they submit an order and the time that the central server acknowledges receipt of this order. In modern electronic limit order trading platforms, the majority of the total latency experienced is due to the time taken for the data to navigate the global telecommunications network between the traders’ office and the central server.

11 This often happens when traders attempt to predict that there is sufficient hidden liquidity in the market to fill a large market order. Sometimes they are correct, and their market order gets matched. Other times they are incorrect and the size of their limit order exceeds the quantity available in the market.
distinguish between a “Fill or Kill” (FOK) or an “Execute and Eliminate” (ENE) status. If the order cannot be fully matched at the time of submission, this status is checked by the matching algorithm. If it is FOK, then the whole market order is cancelled; if it is ENE, as much of the order as can be matched is matched and only the remainder of the order is cancelled.

- In the Paris Bourse (which no longer exists as an exchange – it merged with the exchanges in Amsterdam and Brussels in 2000 to form Euronext) if a market order was submitted at time $t$ then it was only executed against the best price in the limit order book $L(t)$. Any excess quantity in the market order that was not executed against that price was automatically converted into a limit order at that price and placed in the limit order book.

V. EMPIRICAL OBSERVATIONS IN LIMIT ORDER MARKETS

While fundamental understanding of limit order trading is still in its infancy [31], a significant amount of research has been conducted on the statistical properties of observable phenomena in the limit order book (e.g. [6, 9, 33]). Interestingly, many published papers in this area draw contradictory conclusions, and it is not known whether such discrepancies have arisen due to the fact that different markets behave differently or because of methodological variations between papers.

Limit Order Placement

Despite some models assuming the distribution of prices of incoming limit orders to be uniform [37], Bouchaud et al. [9] found that new limit order prices were approximately power-law distributed around the current bid/ask in the Paris Bourse. The authors drew this conclusion based on the plot shown in Figure 2.

A similar heavy (slowly decaying) tail has also been found for equities traded on NASDAQ [33] and the London Stock Exchange [39] via similar methods. Bouchaud et al. claimed that these findings demonstrate how some market participants believe large jumps in stock prices are possible, as traders place limit orders as far as 50% above or below the current bid/ask. All of the markets examined in the aforementioned empirical studies use the price-time priority mechanism, so the observation provides evidence for the idea that patient traders place limit orders far away from the current bid/ask in order to take advantage of any large fluctuations in price. Interestingly, however, the authors cited report significantly different values of the exponent in the observed power laws. Bouchaud et al. [9] suggested a value of approximately 0.6, while Zovko and Farmer [39] gave a value of approximately 1.5. Zovko and Farmer offered a number of possible explanations for the discrepancy: they suggested that sampling frequency might have played a role (we explore this concept in
Section VIII); that their time horizon was not long enough to overcome the role of long-range correlations; that the sample of data they analysed was somehow not representative of typical behaviours as a whole; or that cultural differences might have existed between the markets studied.

Recent theoretical work [12] has cast significant doubt on the integrity of many of the power-law fits drawn from empirical studies in all aspects of science, and it should be noted that none of the power laws reported above are accompanied by formal goodness-of-fit tests (or by likelihood ratio tests for comparisons with other candidate distributions), so one should be cautious about concluding that such a distribution follows a power law. Nevertheless, the distribution’s heavy tail is present in the empirical plots, so the inferences about large swings which are drawn from this seem well supported.

The concept of different traders operating on different time horizons is a common one. Foucault et al. [18] argued that the popularity of the limit order book is due in part to its ability to allow impatient investors to demand “immediacy” (by placing a market order), while simultaneously allowing patient investors to supply “immediacy” to those who will require it at some point in the future (by placing a limit order). The fact that different traders submit different types of order thereby provides evidence for the argument that different market participants value execution speed differently. Foucault et al. [18] identified arbitragers, technical traders, and indexers as being most likely to place market orders (due to the fast-paced nature of their work) and portfolio managers as most likely to place limit orders (because their strategies are generally more focussed on the long term). The theory of how patient and impatient traders affect the market has itself been explored extensively by Zovko and Farmer [39].

Bouchaud et al. [9] observed that limit order arrivals varied significantly in volume, and that the logarithm of the size of incoming limit orders followed a uniform distribution. In terms of the number of incoming limit orders, Bouchaud et al. [9], Challet and Stinchcombe [11], and Biais et al. [6] all noted that incoming limit order flow is greatest nearest to the current bid/ask. Despite this, these limit orders are precisely the ones that are first matched to incoming market orders, so it is not immediately clear what the average shape of the limit order book will be. That is, is is difficult to discern the average quantity available at the best, second best, third best, etc. prices on each side of the limit order book. Perhaps surprisingly, Biais et al. [6] and Bouchaud et al. [9] found the average shape of the order book for the equities studied to be monotonically increasing, away from the best bid/ask. Furthermore, both parties found that the order book had a symmetric shape on average – that is, there was no significant difference (up to sign) between the shapes of the bid and the offer sides of the limit order book. Figure 3 shows a typical limit order book with these properties.

In a recent study [8], Bouchaud et al. tracked the average quantities available at prices deeper in the limit order book (previous studies had considered only the five best prices on each side). They found that monotonicity did not persist beyond the five best prices available in the limit order book. Indeed, the absolute value of the average quantity available for the shares that they studied (Vivendi, Total, and France Telecom) decreased monotonically beyond the maximum observed at the fifth best price. They also noted that the time average of quantities available in the limit order book was very different from the quantities available at

---

In other words, the size of individual limit orders is ignored.
a typical moment in time, as at any given moment the limit order book is generally very sparse (as discussed in Section IV). Roșu [35] conjectured that a “hump” existed in all markets for which large market orders are sufficiently likely, as such a hump represents the tradeoff between the optimism that a limit order placed away from the best bid/ask might be matched to a market order (at a significant profit for the limit order holder) and the pessimism that placing limit orders too far away from the current bid/ask is a waste of time (or, indeed, money, if transaction costs are involved) due to their extremely low fill probabilities. To complicate matters further, Hollifield et al. also noted the existence of a “hump” in the average quantity available for the Ericsson stock (from the Stockholm Stock Exchange) but noted that the maximum occurred at the second best price for both buy and sell limit orders, rather than the fifth. Smith et al. [37] noted that no “hump” is observed when incoming limit orders are assumed to be uniformly distributed over a suitable range of prices, but that a “hump” emerges when the model is refined to reflect the power law like distribution of the price of incoming orders observed empirically. Although their observations are purely empirical (and they are thus unable to offer any insight into what is driving such behaviour), Bouchaud and Potters [9] commented how the shape of the average order book reflects a dynamic competition between the placement of limit orders and the price dynamics that remove the orders close to the current price as they become matched. They propose that such a competition would lead to a universal shape in average limit order books, and suggest that the average shape would be qualitatively similar in many different markets.

Although they still reported symmetry between the bid and the ask sides of the limit order book, Potters and Bouchaud [33] found that the maximum volume of Standard and Poor’s Depositary Receipts (SPY) occurred at \( b(t) \) and \( a(t) \) and decreased monotonically with distance away from these best prices. It is not clear why this should be so – the authors offered the potential explanation that the market flows to which they had access were only a subset of the total market flows. Of course, SPY is not itself an equity in the traditional sense, so people’s valuation or trading strategies for it might differ from that of traditional equities. It will be interesting to see whether future research will provide some concrete explanation for this surprising phenomenon, or indeed reveal it merely to be an artifact of imperfect sampling in some of these empirical studies.

Bouchaud and Potters [9] also conducted an investigation into the quantity available at only the best bid/ask (i.e., \( N(b(t), t) \) and \( N(a(t), t) \)). They found the quantity available at both the best bid and the best ask to follow the same Gamma distribution:

\[
f_X(x) \propto x^{\gamma - 1} \exp \left( -\frac{x}{x_0} \right),
\]

where \( \gamma \) and \( x_0 \) are the shape and scale parameters respectively. \( \gamma \) was noted to vary between 0.7 and 0.8 (a remarkably small range) for the stocks studied. Any Gamma distribution with \( \gamma \leq 1 \) has its maximum at \( x = 0 \), implying that the most probable value of the depth profile at the best bid/ask is very small, which again disagrees with the behaviour found by Potters and Bouchaud for SPY [33].

**Limit Order Cancellations**

Limit order cancellations have also received a great deal of attention recently. It has been noted by Hasbrouck and Saar [22] that about 25% of limit orders on the Island Electronic Communications Network (ECN) – which is part of NASDAQ – are cancelled within two seconds of being placed, increasing to 40% within 10 seconds. Challet and Stinchcombe [11] observed that as many as 80% of the limit orders for the equities that they studied (Cisco, Dell, Microsoft, and Worldcom) ended in cancellation rather than fulfillment. Potters and Bouchaud [33] noted that the observed rate of cancellations monotonically decreases away from the current bid/ask. All of these studies suggest that traders view cancellations as an integral part of their strategy when dealing with limit order markets. As we discuss in Section VII, a failure to acknowledge this fact has led numerous authors to produce models that poorly reflect the real mechanics of limit order trading.

---

13 This is an example of how the observed limit order book \( L(t) \) is an incomplete sample of market participants’ true intention to trade, as discussed in Section IV.

14 SPY is a mutual fund that effectively allows market participants to buy and sell shares in all of the 500 largest stocks traded in the United States.

15 With relevant scaling to account for the fact that there are different volumes available at different distances from the bid/ask.
Conditional Probabilities

With access to a complete time series of all orders and cancellations arriving in the market, it is possible to examine the relative frequencies at which certain types of event happen, given that some criteria has been met. Examples of such criteria could include an increase in the arrival rate of limit orders, a widening of the spread, a reduction in the bid, etc. Biais et al. [6] examined numerous such possibilities and presented the following findings:

1. Market participants are more likely to place limit orders when the spread is large or when there are few limit orders already in the limit order book.

2. Market participants are more likely to place market orders when the spread is small.

3. The conditional frequency of any type of action,\(^{16}\) given that the same action had just occurred, was observed to be higher than the unconditional frequency of that action.

The authors argued that when the spread was small, the market was more liquid. Hence, it was relatively less expensive for market participants to demand immediate liquidity, and so more market orders were placed. Such an argument is in agreement with Foucault et al. [18] whose model found that, the average spread is inversely related to the market order arrival rate.

In order to gain time priority, it is beneficial for market participants to place their limit orders quickly after the spread widens, which could explain the surge of limit orders observed empirically after a widening spread by Biais and Spatt [6]. Taking a zero intelligence approach offers an equally plausible explanation: While the spread is larger, if limit orders are placed at random then it is more likely that an incoming limit order will reside between the greatest rate at and between the effective market order (as discussed in Section IV), which could also explain the higher number of recorded market orders in this situation. Having said this, Biais and Spatt noted that the order placements occur at the greatest rate at and between the spread widens, which could reflect the market's adjustment in valuation due to the information provided by the very existence the trade. They also explored ways to test for mimicking and order splitting in the data.

Another form of clustering that has been observed in the Paris Bourse [6] is comovement of the bid and the ask – that is, when the bid moves down (ask moves up), the ask is more likely to also move down (bid is more likely to also move up). The authors suggested that such behaviour could be due to market participants reacting to information – either some external information has caused a fundamental revaluation of the underlying asset, or (perhaps more reasonably) the downward movement of the bid (upward movement of the ask) is interpreted by other market participants as being information, which caused them to adjust their valuation of the ask (bid) accordingly. Indeed, Potters and Bouchaud [33] noted that the impact of trading on the price is quasi-permanent. In other words, each new trade – by its very existence is interpreted by the market as being a piece of new information. Given this new information, the market adopts the new values of \(a(t)\) and \(b(t)\) and the flow of incoming orders dynamically adapts to this change.

Zovko and Farmer [39] examined clustering of what they define as the relative limit price – that is, the difference between the price at which a sell limit order is placed at time \(t\) and \(a(t)\) (or the difference between \(b(t)\) and the price at which a buy limit order is placed at time \(t\)). By comparing an ordered series of new limit orders with a shuffled version of the same series, they concluded that there was more clustering of subsequent values of relative limit price in the temporally ordered version of the data than there was in the randomly reordered one. The subsequently examined the autocorrelation function of the time series. They found that the

\(^{16}\) The authors defined different “actions” depending on whether new orders were limit or market orders, whether they were buy or sell orders, and also whether they were large or small in size.
autocorrelation function decayed asymptotically as a power law, both in “tick time” (where each subsequent event is given a time index of precisely one greater than the previous event, ignoring the actual time between successive events) and in real time. Unlike many of the other power laws discussed in the literature, this power law is shown to be significant at the 97.5% level. The authors later used this observation as evidence to argue that the price at which incoming limit orders are placed (relative to the current bid/ask) is quite persistent, and that there is no characteristic time scale to the process. This finding supports Potters and Bouchaud’s idea of quasi-permanent price changes: Even when the bid (ask) price changes, the distribution of the distance from the bid (ask) price at which new buy (sell) limit orders arrive remains the same.

Zovko and Farmer also compared the (tick time) relative limit price series with a time series of market volatility, and examined the cross correlation. By using a technique [34] that overcomes the difficulties associated with the examination of non-stationary random variables, the authors were able to test, and reject at the 97.5% level, the null hypothesis that relative limit price and volatility are uncorrelated. Moreover, they found that changes in volatility tended to immediately precede changes in relative limit price, although they were careful to note that it was not clear from their tests whether this was because of some underlying feature that first affected volatility and then affected relative limit price, or because a change in volatility somehow directly caused a change in the relative limit price shortly thereafter.

Wyart et al. [38] have conjectured theoretical reasoning as to why such a cross correlation should exist. Smith et al. [37] suggested another alternative entirely, concluding that their model implied that volatility does not depend directly on market participants’ valuations but rather on the “urgency” (i.e., impatience) that they feel for the trade.

**Price Impact Function**

Market participants who handle large volumes of trade need to be particularly careful when working with limit order markets. Consider, for example, a foreign exchange trader who needs to purchase 30 million GBP by selling USD and is faced by the limit order book shown in Figure 4. If the trader immediately buys all 25 million USD available in the market, it will cost him 15.0643 million GBP right away, and he will only have 5 million USD left to buy. However, by completely wiping out the “sell” side of the limit order book, he has sent a strong signal to all other market participants that he is willing to pay a high price in order to complete his trade. Any observant market participant who wishes to submit a new sell limit order soon after this large market order is likely to do so at a high price, having seen limit orders that were placed at a high price get matched to a market order. Thus, the buyer would again have to pay the high price for the remaining 5 million USD that he still needs to buy.

If, on the other hand, the trader initially only submits a market order for 1 million USD, he will pay 0.6023 million GBP for that 1 million USD but without affecting a(t). Any market participant wishing to submit a limit order with reasonably high priority will be forced to submit such an order close to (or even below) a(t). The buyer can then repeatedly purchase small quantities as the limit order book repopulates without

17 The authors reported the result in real time as being less regular, but still qualitatively similar, to those in tick time.
18 If the latter could be proven, it would support the widely-held belief that many traders consider volatility to be an important factor in making the decision of when to place a limit order [39].
getting himself in a position where he has to pay inflated prices.

Of course, the above example is a simplification of how orders are split in reality. In particular, it assumes that there are no other traders who are also trying to buy, thus reducing the control the individual buyer in question has over \( a(t) \). It does, however, illustrate how market participants in a limit order market need to carefully manage their market impact:

**Definition.** The market impact function \( \phi(\omega, \tau, t) \) is the change in \( a(t + \tau) \) (resp. \( b(t + \tau) \)) that is directly caused by a buy (sell) market order of size \( \omega \) that is placed at time \( t \).

Notice that the above definition allows the possibility that price includes some informational content. That is, by submitting a market order at time \( t \), it is possible that the market participant affects the best price in the market not only at time \( t \) but for all times \( t + \tau \), with \( \tau > 0 \). This idea – that market impact is twofold (i.e., consisting of both temporary and permanent components) – is due to the seminal work by Almgren and Chriss [4].

Less general is the concept of the instantaneous price impact function \( \phi(\omega, t) \), which examines only the immediate effect that a market order of size \( \omega \) has on the best bid/ask.\(^{19}\) Over the past two decades, the instantaneous price impact function (when considered as a function of \( \omega \)) has been at the centre of a hotly contested debate in the literature. Early theoretical work on the limit order book by Glosten [19] suggested that there need not be any particular shape to the instantaneous price impact function. A year later (in their empirical study of the Paris Bourse), Biasi et al. [6] published findings of a “weakly concave” average shape of the limit order book, giving rise to an instantaneous price impact function that is convex in \( \omega \). Numerous recent studies – both empirical [32] and theoretical [37] – have found the instantaneous price impact function to be concave in \( \omega \). This concavity seems to be the generally accepted view at present: Daniels et al. [15] noted that for small values of \( \omega \), the study of empirical data which had been conducted seemed to suggest a relationship of the type \( \phi(\omega) \sim \omega^3 \). The authors commented, however, that estimates of the value of \( \beta \) were poor and often contradictory, concluding that their “best guess” equipped with the research conducted up to that point was around \( \beta \sim 0.5 \). The authors go on to derive an estimate for \( \beta \) using a mean-field approximation, and again conclude that \( \beta \sim 0.5 \).\(^{20}\) Smith et al. developed a model of the limit order book based on zero intelligence that – under a mean-field approximation – also predicted such concavity.\(^{21}\) The question of why the original study on the Paris Bourse found a “weakly concave” average shape of the limit order book (and thus an instantaneous price impact function that was convex in \( \omega \)) remains unanswered, although the authors were careful to note that their data did not include any information about hidden orders. It is possible that such hidden orders could have played a significant enough role in the shape of the order book to skew their results.

Challet and Stinchcombe [11] explicitly highlighted that the informational content of trades is a particularly important issue when considering the price impact function, particularly in the light of Potters and Bouchaud’s conclusions (that every trade is itself interpreted by the market as being new information). They suggested that there should always be a clear separation between between what they called “real” and “virtual” market impact. In a real market, the “real” impact of a market order is not only an instantaneous change in the limit order book but also a signal to other market participants that could potentially change the limit order book further in the future. “Virtual” market impact is far easier to quantify – it is the study solely of how the limit order book would change instantaneously if a particular market order were submitted at a time \( t \), ignoring any effects that this might have in the long term. Clearly understanding virtual market impact is far less useful than understanding real market impact, but real market impact is very difficult (if not impossible) to observe. The best approximation to real market impact arises from examining historic data to find times when certain scenarios occurred and noting how the market behaved afterwards. Even then, there is no way of telling how the market “would have” behaved if a given order were never submitted, so real market impact is exceptionally difficult to quantify in practice [37].

---

\(^{19}\) The terminology “price impact function” is used inconsistently in the literature. Many authors use the term “price impact” to discuss the long-term effects of market order submissions whereas others use it when only considering the instantaneous effects. We introduce the “instantaneous” prefix to distinguish between these two possibilities.

\(^{20}\) Interestingly, the authors acknowledged that their mean-field approximation was a poor reflection of real limit order books due to the fact that they let their resolution parameters tend to zero. In a more detailed exploration later in the paper (in which they do not take such a limit), the authors found two different values of \( \beta \), depending on the timescale used. They still found the value \( \beta \sim 0.5 \) for long timescales, but they found that \( \beta \sim -0.5 \) for very short timescales. They were unable to offer any theoretical explanation of why this should have been the case.

\(^{21}\) Notice that concavity of the instantaneous market impact function highlights the fact that (assuming traders are not behaving in an entirely zero intelligence manner) order-splitting decisions are based on more than just the instantaneous effect that they will have on the market. Otherwise, it would always be cheaper for them to submit large orders all at once than it would for them to break them into smaller pieces.
VI. THE IMPACT OF THE LIMIT ORDER BOOK ON MARKETS

Irrespective of whether zero intelligence, perfect rationality, or some hybrid of the two is used by market participants, it is undeniable that the move to limit order trading has fundamentally affected markets. Daniels et al. [15] noted how the basic limit order book function of storing supply and demand from market participants – a function not present in quote driven markets – in and of itself induced structure in market prices. If a market participant bases their trading decisions on what they know about the state of the market, by giving them access to more information about other market participants, the very existence of the limit order book must have changed the market. There is also empirical evidence to support this idea: Boehmer et al. [7] found that when the NYSE made more information about the limit order book available to all market participants, there was a detectable change in order flows. If agents act with zero intelligence, increasing the range of options available to them from simply “buy” and “sell” to also include posting limit orders clearly affects the market for all participants.

Strategies in Limit Order Trading

It is a recurring theme in the literature (see, for example, [6, 18, 37]) that the specific rules and resolution parameters ($\sigma$ and $dp$) of a particular limit order market greatly affect the way that the market’s participants trade. As understanding of the limit order book has grown over time, trading strategies have emerged that exploit these intricate details. We now explore a simple trading strategy that takes advantage of the fact that $dp$ is typically set to some strictly positive value. Consider, for example, a trader who wishes to sell 1 unit of the asset being traded in the relatively sparse limit order book shown in Figure 5. Assuming that the trader wishes to maximise the amount that he receives for the asset, but that he also does not want to sit and wait for a long period of time before his limit order is matched, what should he do? In a market in which the trader is only provided with information about the best price in the market, he would most likely choose to match (or perhaps even beat) the best price ($1.9632$) so that his order is matched relatively quickly. However, presented with the full information of the limit order book, a strategy that is arguably better is to place his limit order at $1.9635$. The trader observes that there is only one limit order in the limit order book at price $1.9632$, and once this order has been matched (if no other limit orders at a better price than his order at $1.9635$ arrive), he will be offering the best price in the market. If, at any time, another market participant submits a limit order at a better price, the trader can simply cancel his limit order at $1.9635$ and match or beat the other order if he sees fit. Until some other limit order arrives, his limit order will be the best in the market. If a buy market order arrives during this time, the trader will match at a much better price than if he had originally placed a limit order to beat the first existing limit order of $1.9632$. Biais et al. [6] found empirical evidence to support the idea that traders do indeed employ such strategies. If numerous traders are simultaneously employing this same technique, “undercutting” will result as they both compete to be ahead in the priority queue.
Mixed Strategies – A Game Theoretic Approach

As discussed in Section IV, latency is an important consideration for market participants, especially those who are battling for time priority in the limit order book. Although latency times can be as low as a few milliseconds, the latency involved in a round-trip for an order travelling from Asia to North America can be as large as 250 milliseconds. Biais et al. [6] presented the following example of why (rational) traders might therefore want to play a mixed strategy when submitting limit orders. Imagine a market in which \( dp = 0.1 \) and in which \( b(t) = 150.7 \), with some large quantity of the asset available at this price in the limit order book \( L(t) \). Three traders each want to submit a single-unit buy limit order that will position them at the front of the priority queue in the market. Trader 1 experiences latency \( \varepsilon_1 \), trader 2 experiences latency \( \varepsilon_2 \), and trader 3 experiences latency \( \varepsilon_3 \), where \( \varepsilon_1 < \varepsilon_2 < \varepsilon_3 \). If they all simultaneously submit buy limit orders at 150.8, neither traders 2 nor trader 3 will achieve their objective of being at the front of the priority queue, as they will receive equal price priority but lower time priority than trader 1. The only way for them to gain priority is to subsequently cancel their initial limit order and to submit a new limit order with an even higher price. Game theory dictates, therefore, that traders should not select the price at which to submit limit orders in an entirely deterministic way if they wish to compete for priority [30]. Otherwise, observant market participants will always be able to predict their actions and to take advantage of them (particularly if these other market participants have lower latency times). Instead, they should include a random element in their initial choice of limit order price, which cannot be manipulated by other people. Sometimes this will give them their desired priority and other times it will not, but it will stop people from being able to predict and manipulate their choices.

VII. MODELLING THE LIMIT ORDER BOOK

Building a useful model of the limit order book has proven to be an exceptionally difficult task [11]. The basic difficulty with modelling the limit order book is striking an acceptable balance between producing a model that is tractable versus producing one that accurately reflects the state of the market. Naturally, the “ideal” state of this balance also depends on the purpose of the model. Particular problems that arise repeatedly are awkward facts such as the arrival processes of new orders not fitting any common distributions. In particular, the clustering of new order arrivals implies that they do not behave as a Poisson processes, which many authors have previously assumed in order to maintain analytic tractability. In contrast, some models attempt to examine the behaviour of traders at the microscopic level in order to study the effect that such behaviours have on markets. These models often run into the difficulties associated with quantifying qualitative traits (see Section IV). Furthermore, the utility of such models is questionable, as there is no guarantee that a particular trader has to act in a particular way when presented with a given set of circumstances. At present, the best compromise between these two extremes seems to be to coarse-grain the microscopic behaviours of market participants and to instead study the impact of the resulting aggregate order flows on the market [37].

Cont and Talreja [14] presented the current “state of play” as falling broadly into three categories. First, there are a collection of tractable models that are based largely on examining trajectories of appropriately defined dynamical systems. Second, there is a collection of purely stochastic models that attempt to fit distributions to the aggregate flows in the limit order book. Finally, there is a collection of models that are designed to accurately mirror the nuances of limit order trading – unsurprisingly, these models invariably end up being analytically intractable, so observations based on Monte Carlo methods are the best they can offer. Moreover, because of the small discrepancies between matching algorithms and rules used in different limit order books, even the models that have been designed to closely replicate the mechanics of limit order trading are not able to perfectly mimic the rules of any particular exchange. Instead, they settle for capturing the essential common features of limit order markets. In this section, we examine several models of the above types. We also chart a short history of such modelling in order to evaluate what contribution (if any) such models have provided in terms of understanding the fundamental mechanics of the limit order book.

Arguably the most important paper in terms of deepening understanding of the limit order book to date is Smith et al.’s “Statistical Theory of the Continuous Double Auction” [37]. While the authors noted that their model does not perfectly represent the true mechanics of limit order trading – in particular, the model assumes indepenence between flows at different price levels and also that incoming limit orders arrive with a price that is uniformly distributed across all of the prices that would make the limit order an effective limit order – the authors were able to analytically probe their model in sufficient detail to make some fundamental observations about the bid-ask spread, market volatility, the shape of the limit order book, the price impact function, and the probability and time to limit order matching. Smith et al. also built a reality check into their work — in addition to deriving models, they deduced some conservation constraints that they argued
must be met by the price process in equilibrium.\textsuperscript{22} The authors were then able to check whether the results of any Monte Carlo simulations that they performed on their model obeyed the conservation constraints. By noting when the conservation laws had been violated, they were able to pinpoint the weaknesses in their model. They frequently cited the coexistence of different time-scales to be problematic (as discussed below). Smith et al. found that their model worked best in situations in which the minimum order size was (relatively) large, but they observed from empirical data that minimum order sizes in real markets were somewhat smaller than those in their optimum range.

A Brief History of Modelling The Limit Order Book

Identifying a fully time-ordered path of progress in modelling the limit order book is a difficult and highly subjective task. In the early stages, work was carried out in both the economics and the physics literature, and Smith et al. \textsuperscript{37} argued that progress in the two fields was largely independent. Over time, authors have come to understand the coupled nature of the limit order book and the order process itself. The order process appears to depend on the state of the limit order book, which itself depends on the state of the order process, and so on. Although Smith et al. tracked the progress in the economics literature back to a paper published in 1982 \textsuperscript{29}, they noted that the first published market model that operated under a set of assumptions that remotely resembled a real limit order market was due to Bak et al. in 1997 \textsuperscript{5}.\textsuperscript{23} By formulating a model based on a collection of repeatedly shuffled prices, Bak et al. were able to explore the market using modelling techniques based on the idea of reaction-diffusion. While this “first step” in modelling the limit order book was an important one, its modelling assumptions poorly reflected the reality of limit order trading, so the present-day use of the model is limited.

Maslov’s “Simple Model of a Limit Order Driven Market” in 2000 \textsuperscript{28} was the first paper to consider a setup that was immediately identifiable as having a limit order book structure. In an attempt to keep his model simple, Maslov assumed that cancellations were not permitted in the market. Hence, as discussed in Section V, his model has limited practical use. Despite this, Maslov was able to demonstrate that the price time series generated by his model (in a zero intelligence framework) displayed some statistical similarities to those qualitatively observed in real limit order markets (including long range correlations, fat-tailed distributions for price changes, and price series with non-trivial Hurst exponents).

Numerous authors have since found techniques from physics useful for deepening understanding of the limit order book. Challet and Stinchcombe \textsuperscript{11} considered the limit order book to be a system of particles on a discrete pricing grid.\textsuperscript{24} The authors were then able to draw analogies with evaporation or annihilation of such particles in order to perform an analysis of the dynamics of the system. In particular, the authors note that under such a setup there are two types of particles (limit orders and market orders) which can be placed on the lattice. If two particles of opposite type are placed at the same position on the lattice, they annihilate each other — such an annihilation corresponds to a market matching. In addition, limit order particles can also “evaporate,” which corresponds to the limit order being cancelled. Using their model, the authors were able to estimate the limit order lifetime distribution (the distribution of the length of time between when a limit order is placed and when it is either matched or cancelled) and the shape of the distribution of incoming orders, and, crucially, to acknowledge the key role that order cancellations play in the evolution of prices.

Another technique from physics which has been used for researching the limit order book is that of mean-field approximation. A mean-field approximation is when a many-body system with complex interactions is approximated by replacing all interactions to any one body with some average interaction. In the context of Challet and Stinchcombe’s model, a mean-field approximation could be made by approximating the behaviour of each individual particle to be the average behaviour across all appropriate particles in the system. Slanina \textsuperscript{36} explored the distribution of price changes by making two mean-field approximations. First, rather than considering the actual prices of incoming limit orders, the prices were assumed to be uniformly distributed across all prices. Second, behaviour on both sides of the limit order book was assumed to always be symmetric. With this setup, one can derive a closed-form power-law expression for the distribution of price changes.

\textsuperscript{22} An example of such a conservation constraint is the idea that if the number of orders in the limit order book is bounded, then the long-term arrival rate of new limit orders must equal the sum of the long-term cancellation rate of limit orders and the long-term arrival rate of market orders.

\textsuperscript{23} Although this equivalence is not immediately obvious, if all traders in Bak et al.’s model are deemed to submit their “estimated price” to the market in the form of a limit order and subsequently modify these limit orders as they update their estimations, then the equivalence to a limit order market becomes clear.

\textsuperscript{24} Although not the first to publish such a representation of the limit order book in this way (e.g., \textsuperscript{5} also does this), Challet and Stinchcombe’s was the first model that attempted to mimic the “real” order flows of a limit order market, including the flow of cancellations.
Slanina showed that his derived distribution exhibited negative autocorrelation at high frequency, which is in line with what has been empirically observed in the price time series of equities [16].

Smith et al. [37] also made use of mean-field approximations to derive closed-form expressions about the limit order book. They also used dimensional analysis to reduce the complexity of their model by demonstrating that there are really only two free parameters, rather than the apparent five. This led them to conclude that minimum order size plays a significantly stronger role in affecting the model's behaviour than does tick size. They used this reasoning to justify working in the limit $dp \to 0$, but were careful to note that this was not a perfect representation of a real market.

**Game-Theoretic Approaches**

The underlying spirit game-theoretic approaches to modelling the limit order book is based on what Hollifield et al. [23] called a “structural estimation of auction models.” In other words, some functional form (either deterministic or stochastic) of private valuation of the asset being traded is assumed, and trades are assumed to occur by individuals who wish to maximise some utility function. The difficulty in such an approach arises in attempting to quantify agents’ behaviour in some mathematical manner. Hollifield et al. attempted to do so, but the model that their assumptions led to yielded inaccurate predictions of market behaviour (for example, in their model, limit order placement would not have such heavy tails as those observed in empirical studies, because the expected payoffs from submitting such limit orders are too low to justify such an action). Hollifield et al. formulated the model as a generalisation of Foucault’s earlier work [17], in which he proposed a model in which valuations were based upon independent and identically distributed random variables drawn from fixed distributions. Hollifield et al.’s extension of this work allowed these distributions to depend on quantifiable properties of the market that evolve with time. While they acknowledged that ideally they would have conditioned on the entire history of the limit order book, instead (in order to ensure tractability) they restricted their conditioning to previous order quantities, the quantity available at the bid/ask, the (lagged) volume of trading in the market, volatility, and the present time of day. Although this choice produced a tractable model, the authors rejected their model in their conclusions due to the fact that it forced them to make deductions that were not in line with their empirical observations and “common sense.” Consequently, the authors concluded that their underlying assumptions (and the applicability of the model) could only be considered acceptable for traders with extreme private valuations of the asset being traded.

Foucault and Kandel [18] attempted to understand some of the empirically observed phenomena in a market by assuming that market participants arrive in a specified way: First a buyer arrives and places a buy order of some kind; then a seller arrives and places a sell order of some kind. Along with two other assumptions (that submitted limit orders cannot be cancelled and that limit orders can only be placed within the current bid-ask spread), this allowed the authors to explain certain behaviours such as how the average time until a new order arrived, conditional on the spread at the time of the arrival of the previous order, increased with the size of the spread. Later in the paper, the authors asserted that some of their assumptions could be relaxed without fundamentally altering the model, but the assumption of the alternating arrival of buyers and sellers is necessary in order for the proofs in the paper to hold. Therefore, the paper’s conclusions must be treated with caution. Nevertheless, the authors presented an interesting quantification of the idea of “rationality” in the specific situation in which traders wish to maximise their profits but experience some “time penalty.” The model assumed that the market could be represented by a discrete pricing grid $\{1, \ldots, n\}$, where the distance between adjacent points on the grid represented $dp$ in the real market. Then, for $j \in \{0, \ldots, n-1\}$, a $j$ limit order is a limit order that results in the spread $s(t)$ being equal to $j$ immediately after it is placed. Writing $T(j)$ for the expected time until a $j$ limit order is matched to a market order and writing $\delta_i$ for the waiting cost experienced by market participant $i$, then the expected waiting cost of a $j$ limit order is $\delta_i T(j)$. The authors then presented an argument that each trader $i$, observing spread $s$, chooses their optimal order

---

25 Portfolio managers often unwind any positions that they have put themselves into before the end of the trading day in order to minimise their exposure to market movements overnight, so their trading behaviour in the afternoon often differs greatly from that of the morning.

26 However, the conditions that must be met in order for the assumptions to be relaxed seem to me to be unrealistic.

27 Notice that the authors assume that $T(j)$ does not depend on the price that the limit order is placed at, but rather only on the spread after it is placed.

28 This clearly varies between market participants: “impatient” traders have a relatively higher waiting cost per unit time than do “patient” traders.
$\pi^*_i$ among orders $\pi_i(j)$ by solving:

$$\pi^*_i = \max_{j \in \{0, \ldots, s-1\}} \pi_i(j) = jdp - \delta_i T(j)$$

This allowed the authors to make the precise definition:

**Definition.** An order placement strategy for trader $i$ is a mapping $o_i(\cdot)$ that assigns a $j$ limit order, $j \in \{0, \ldots, s\}$, for every possible spread $s$.

If it is assumed that there are only two possible values that $i$ can take (the authors deemed these “patient” and “impatient”, and denoted them $P$ and $I$, respectively), then one can make the following definition:

**Definition.** An equilibrium of the trading game is a pair of order placement strategies $o^*_p(\cdot)$ and $o^*_i(\cdot)$ such that the orders prescribed by the strategies solve equation (3) when the expected waiting time $T^*(\cdot)$ is computed assuming that traders follow strategies $o^*_p(\cdot)$ and $o^*_i(\cdot)$.

The authors used their results to offer an interpretation of Biais et al.’s empirical study [6], claiming that the observations documented in [6] are exactly what their model would predict if all traders in the market were homogeneous – that is, if all traders in the market would make the same decision when confronted with the same state of the limit order book.

Roșu [35] also used a game theoretic approach to model the limit order book. Beginning with a setup similar to Foucault and Kandel [18], Roșu made the additional assumption that market participants could dynamically modify their limit orders by making use of cancellations. Interestingly, rather than complicating the model further, this assumption greatly simplified the model because it now meant that in equilibrium, all market participants with limit orders of the same type in the limit order book must have equal expected utilities – otherwise they would modify their limit order accordingly. Roșu went on to derive a market equilibrium after making some large modelling assumptions, e.g., the limit order book having a finite capacity. Roșu’s model included unobservable parameters such as the “patience coefficient” and a “competition parameter,” and consequently had limited scope for practical application, but the theoretical insight into utility provided by the model is considerable.

There has recently been a notable absence of game-theoretic explorations of the limit order book in the published literature, perhaps due in part to the difficulty of forming tractable analytic models when working with concepts such as perfect rationality, information, and utility.

**Dynamical Systems Approaches**

Luckock [27] took an interesting approach to modelling the limit order book. First, assuming that $dp \rightarrow 0$ so that the equations he originally derived in a discrete setting became differential equations. He was then able to solve this system of equations in order to find the steady-state distribution of the best bid and ask prices. He was also able to infer that the stationary distributions of the best bid and ask prices were determined solely by the elasticity of supply and the elasticity of demand, and also to derive some simple trading strategies. As all this work was based on the assumption that $dp \rightarrow 0$, Luckock’s observations have limited application to real trading.\textsuperscript{29} Indeed, Luckock acknowledged the fact that the strength of his work was in the simplicity and tractability of the model, not in its ability to accurately replicate market dynamics.

\textsuperscript{29} As discussed in Section VI, markets actually have some fixed $dp > 0$, and market behaviour is noted by numerous authors (eg. [6, 18]) to be dependent on the size of $dp$. 

Markovian Approaches

Among the models in the existing literature, the one that seems strike the best balance between analytic tractability and realism is that of Cont and Talreja [14]. By making the same mapping onto a discrete pricing lattice \( \{1, \ldots, n\} \) as was made by Foucault and Kandel [18], the authors observed that the state of the limit order book could be tracked through time as a continuous time Markov process \( X(t) := (X_1(t), \ldots, X_n(t))_{t \geq 0} \), where \(-X_p(t)\) is the number of limit sell orders and \(X_p(t)\) is the number of limit buy orders present in the limit order book at time \(t\), for \(p \in \{1, \ldots, n\}\).\(^{30}\) Under this setup, it follows immediately that

\[
\begin{align*}
 b(t) &= \sup \{1, \ldots, n | X_p(t) < 0\} \quad \text{and} \quad a(t) = \inf \{1, \ldots, n | X_p(t) > 0\}
\end{align*}
\]

The only possible transitions that \(X(t)\) can undergo stem from the arrival of a new limit order at a price \(n\), the cancellation of an existing limit order at a price \(n\), or the arrival of a new market order.

To construct a tractable model, the authors made various assumptions about the rates at which these events occurred. They assumed that arrivals of limit orders at price \(p\), for \(p \in \{1, \ldots, n\}\), occur in unitary size, and as a Poisson process at rate \(\lambda(p)\). They assumed that cancellations of limit orders at price \(p\), for \(p \in \{1, \ldots, n\}\), occurred as a Poisson process at rate \(\theta(p)\) for each existing limit order. Hence, they assumed that cancellations of limit orders at price \(p\) occurred at the rate \(|X_p|\theta(p)\). They assumed that sell market orders and buy market orders each arrived as a Poisson process at rate \(\mu\). Additionally, they assumed that each of these Poisson processes were mutually independent, across both different prices and different types (so, for example, the incoming limit orders at price \(3\) are independent of the incoming limit orders at price \(4\) and independent of all cancellations and of all market orders).

Under this setup, \(X(t)\) evolves as a continuous time Markov process (on state space \(\mathbb{Z}^n\)). Then, using standard results about such processes, the authors proved two results. The first result concerns admissible states of the limit order book (that is, when \(a(t) > b(t)\)):\(^{31}\)

**Theorem 1.** If \(X(0)\) is admissible, then \(X(t)\) is admissible almost surely \( \forall t \geq 0 \).

The second result concerns ergodicity:

**Theorem 2.** If \(\theta := \min \{\theta(i)| i \in \{1, \ldots, n\}\} > 0\), then \(X\) is an ergodic Markov process with a stationary distribution \(\pi(N_1, \ldots, N_n) = \mathbb{P}(X_1 = N_1, \ldots, X_n = N_n)\).

Ergodicity is an exceptionally useful property, because is implies that expectations of the form \(\mathbb{E}[f(X_\infty)]\) can be estimated by simulating the limit order book over a large time horizon and averaging \(f(X(t))\) over the simulated path. That is,

\[
\frac{1}{T} \int_0^T f(X(t))dt \to \mathbb{E}[f(X_\infty)] \quad \text{almost surely as} \quad T \to \infty.
\]

The authors then compared the results obtained using analytic techniques based on their model; using “semi-analytic” techniques based on their model via a numerical inversion of the Laplace transform; using Monte Carlo simulations from their model; and using actual data. They found (across all of the above methods) that the long-term behaviour and the volatility of their model fitted with the empirical observations of real data.

An interesting aspect of the model in [14] is that it allows direct calculations of conditional probabilities of certain events (such as the probability of successfully executing a limit order before the mid-price moves, and the probability of successfully matching both of a pair of limit orders – one at \(b(t)\) and the other at \(a(t)\) – and thus earning an amount equal to \(s(t)\)). These findings are directly relevant to traders as an application of high-frequency trading and provide insight into the merits of placing market or limit orders, given a particular set of market conditions.

The authors also presented a strategy for a statistical arbitrage,\(^{32}\) which was shown in the paper to be mildly successful but had a sufficiently small profit that it was deemed unlikely to survive transaction costs.

---

30 That is, the authors adopted the convention that limit buy orders be counted as being negative in their size.
31 More formally, a state \(x \in \mathbb{Z}\) is admissible if and only if there exists \(k, l \in \mathbb{Z}\), with \(1 \leq k, l \leq n\), such that: \(x_p \geq 0\) for all \(p \geq l\), \(x_p \leq 0\) for all \(p \leq k\), and \(x_p = 0\) for all \(k < p < l\).
32 A statistical arbitrage is a trade (or sequence of trades) with a positive expected return and a bounded loss. It is much weaker than a true arbitrage, but it is still a desirable thing to find as it should be profitable (on average) in the long run.
In this survey, we have examined how recent work has contributed to a deeper understanding of the evolution of price process in limit order trading. Although a considerable number of authors have made significant progress in the area over the last two decades, general understanding of the limit order book – particularly in its modelling – is still in its infancy. Many authors (e.g., Bouchaud et al. [8]) have noted the inadequacy of considering order flows as being independent of the state of the limit order book, and while numerous authors (e.g., [18, 35]) have been able to create models that predict the empirically observable effect of order flow on the limit order book, only the most recent work (e.g. [14, 23]) has been able to also acknowledge the effect that the state of the limit order book has on order flow.

The flavour of much of the early published literature on modelling of the limit order book was to create a model, test (either analytically or via Monte Carlo simulation) that it displayed the same statistical properties as actual data, and then to attempt to derive some simple trading strategy based on implications from the model. Aside from doing relatively little to deepen understanding of the limit order book (aside, perhaps, from highlighting those modelling assumptions that lead to unreasonable conclusions), the value of discussing such trading strategies is also questionable due to the inherent difficulties involved in testing them. It is an unavoidable fact that every new limit order changes the landscape of the limit order book, and Potters and Bouchaud [33] noted that the impact of every newly arriving order on the price extends well beyond the lifetime of the individual order. Although there has recently been some progress [24] in developing methods to theoretically compare the performance of candidate trading algorithms, the only way to decide whether a trading strategy has any real commercial value is to test it in a real market. Theoretical speculation about the performance of a trading strategy has little value.

Equipped with recent, high quality financial data, the research community is now in an excellent position to explore the intricacies of limit order trading at great depth. An issue of paramount importance is the apparent notion that limit order book data which is sampled at a very high frequency (e.g., on a tick-by-tick or many-times-per-second basis) exhibits very different properties than data sampled at lower frequencies. The timescales between successive events in the limit order book are often of the order of milliseconds, and as higher frequency data has become available, many empirical observations which were previously accepted have again come to be challenged. Examples of this are serial autocorrelation in returns (which appear absent in a second-by-second analysis, but are certainly present in a tick-by-tick analysis of the market and – according to Smith et al. [37] – persist in timescales up to the order of 50 average market order inter-arrival times); the variation of volatility with trading activity (which appears to follow a positive relation on a second-by-second analysis, but has been found to follow an inverse relationship by Wyart et al. [38] at higher frequencies); differences in the parameter values for the price impact function reported by Daniels et al. [15] at high and low frequencies; and Smith et al.’s [37] observation that “diffusion” (i.e., the rate of change of volatility) occurs faster at shorter timescales than it does at longer ones. Previous publications have been vague about the exact particulars of their data, but moving forward we firmly believe that it will become imperative for authors to disclose as much information about data sampling as their sources allow. This is obviously a difficult issue when much of the data in the field is commercially sensitive, but publishing as much information about the nature of their data as possible should be a primary consideration for researchers in the future. Only then will some of the mysteries based around different authors drawing contradictory conclusions about the same relationships have any hope of being resolved.

At present, it remains unclear whether alternative approaches to limit order book modelling really are “different,” or are actually just different manifestations of the same ideas. It may later be proven, for example, that a dynamical systems approach is precisely the same as a Markovian queueing approach under some suitable limiting regime. No such equivalence is immediately clear from the existing literature, but it will be very interesting to see whether, as modelling techniques advance, different approaches end up yielding identical results.

A key unresolved question is the relative applicability of perfect rationality and zero intelligence assumptions. Bouchaud et al. [8] state that although information can be observed to play some role in forming prices, for many purposes this role is secondary to market structure. Zovko and Farmer [39] also comment on how striking regularities can emerge when different human beings are faced with the same decision problem – much like limit order trading. While undoubtedly a difficult task to complete – especially in the often highly secretive world of financial research – compiling some conclusive results about how large numbers of market participants react in a variety of situations would greatly aid understanding of order flows. An interesting proposal which has not yet been thoroughly explored is the idea that individual traders may employ different trading strategies which, across different traders, average to something resembling zero intelligence. In other words, it might be possible that individual “intelligences” somehow cancel each other out. This is a particularly difficult inference problem due to the enormous number of interacting traders involved. Indeed, Parlour and Seppi [31] stated that understanding how individual trading strategies might aggregate to the order flows observable in data was one of the most important problems facing researchers in the future.
By giving traders the freedom to evaluate their own need for immediate liquidity in the market, the limit order book has revolutionised the process of trading. In addition to evaluating their buying and selling needs, market participants must now also consider the tradeoff between having their order matched quickly and getting the best possible price for their trades. Entirely new trading tactics have grown from the rich structure of the limit order book. Many researchers (e.g., [6, 30]) have focussed on how certain limit order trading strategies affect the price process, while others (e.g., [20, 37]) believe that it is the fundamental economic need to store supply and demand that plays the primary role in the price formation, not strategic thinking. As financial practitioners become increasingly accustomed to limit order trading and as researchers begin to better understand the complex mechanisms underpinning the process, it will be fascinating to observe which old problems will be solved – and, indeed, which new obstacles will be presented – during future investigation of the limit order book.

Acknowledgements

We would like to thank Jean-Philippe Bouchaud for his permission to reproduce Figure 2, which originally comes from [9].


[26] E. Smith, J.D. Farmer, L. Gillemot, and S. Krishnamurthy, Statistical theory of the continuous double auction,
Quantitative Finance 3 (2003), no. 6, 481–514.
