Introduction to Limit Order Book Markets
NUS@Oxford 2019

Ben Hambly
Mathematical Institute

June 12, 2019

Introduction

In this lecture, we review the basics of *Limit Order Books* (LOBs). These are the electronic trading platforms now used by financial markets worldwide.

We will see how they operate and how they differ from traditional markets.

We will also look at some simple ideas for modelling them. Other lectures will look at strategies for trading in such markets.

Types of market

The simplest way of trading is an individual transaction between a buyer and a seller, perhaps with the help of an agent. Example: buy a house. This is, roughly, the over-the-counter (OTC) market. In finance it's only used for one-off items or complex contracts.

Standardized items such as shares, bonds, commodities etc are mostly traded centrally.

- Quote-driven markets (also broker-dealer);
- Open outcry markets;
- Limit Order Book (LOB) platforms.

Quote-driven markets

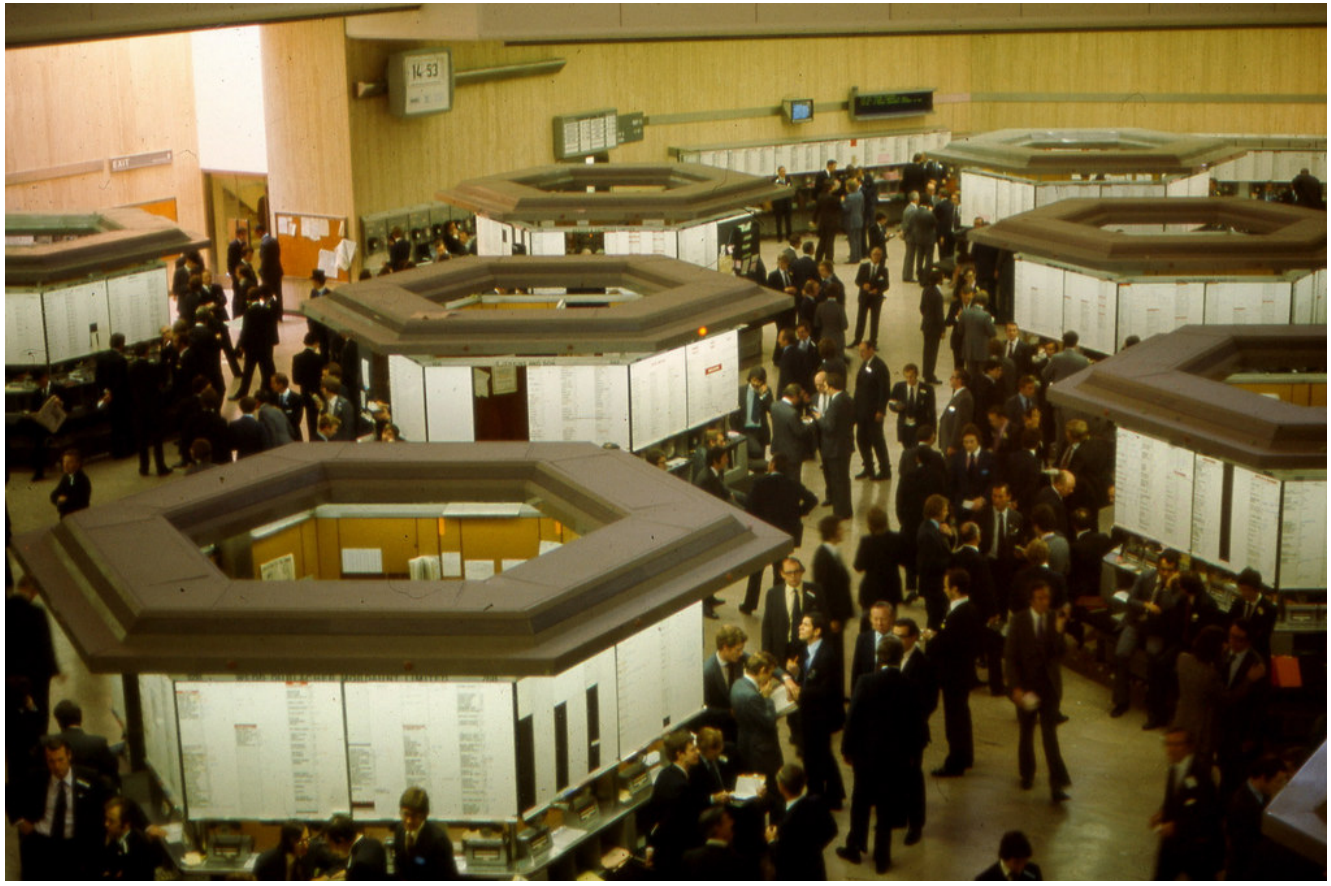
Dealers, also called *market-makers,* buy and sell stock from clients, usually via a *broker*. The dealers try to buy low and sell high. They post *quotes* indicating their prices for buying and selling.

- Dealers have to have capital and access to the market.
- Dealers have to hold the stock (risky).
- Not transparent, can be uncompetitive.
- Clients can trade whenever they like (they can be patient or impatient) but dealers have to be patient and wait for trades.
- Tick sizes[1] can be large (large transaction costs).

[1] The *tick size* is the smallest price division used.

- We see the FX market with dealers at the airport offering to change your money. Note the difference between the price to buy and to sell a given currency.
- They make profit from crossing the bid-ask spread
- Market Makers offer a crucial service: provision of liquidity

https://www1.oanda.com/currency/live-exchange-rates/



The floor of the London Stock Exchange in 1975

Photo: David Buckley on flickr

Open outcry markets

Buyers and sellers all meet and competitively advertise their prices. This is usually in a space called a *pit*. It is highly skilled and technically difficult. Most trade is by hand signals.

- Cuts out the dealers (no need to hold inventory).
- Requires very specialised skills.
- Open to dispute (disputes are surprisingly rare).
- Relatively slow to operate.
- Not very transparent (eg there is a delay between a trade being made and the information being posted for the market to see).
- Clients need a trader to act for them (transaction cost!).

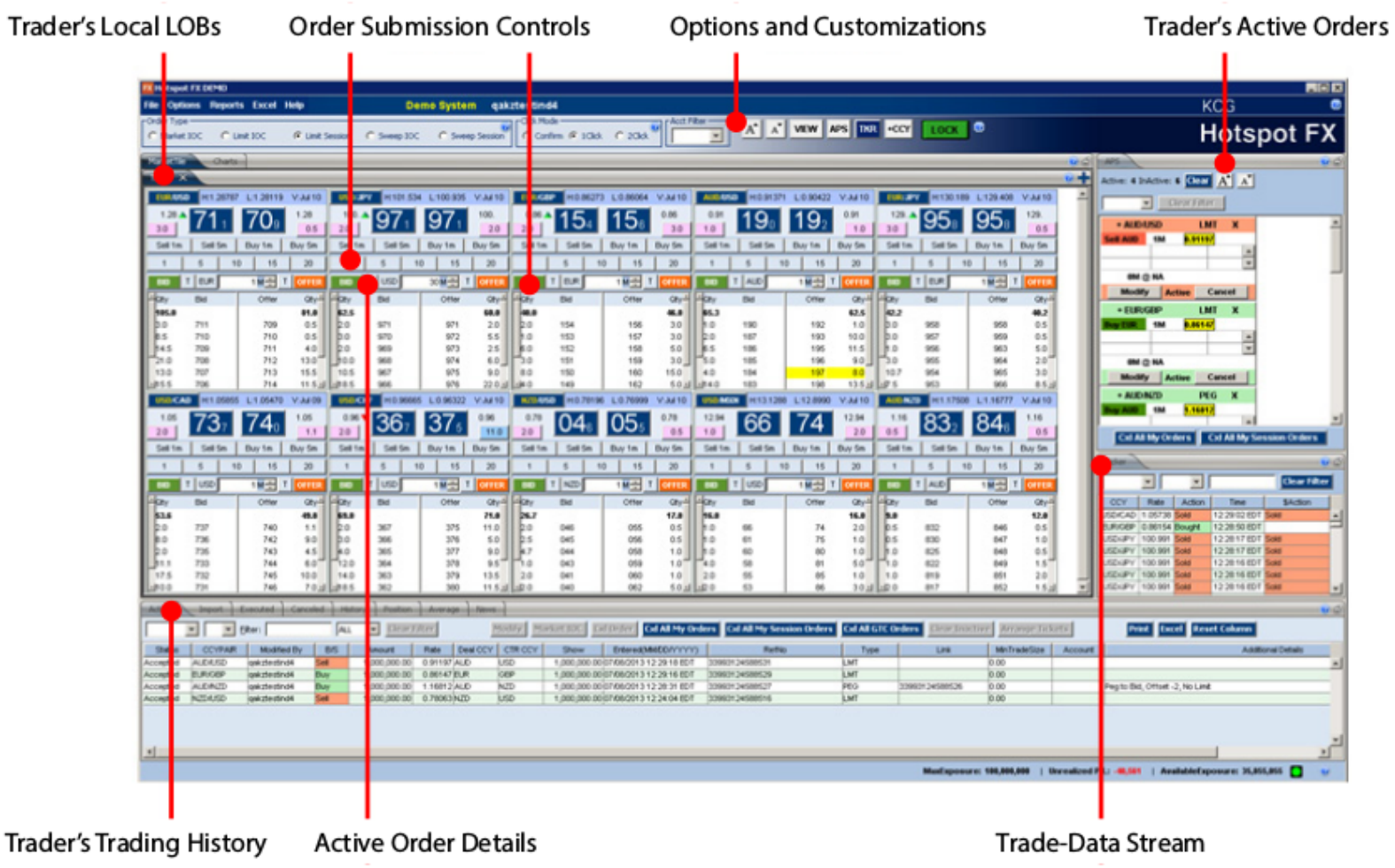
Trading on the CME floor.
Photo: www.chicagobusiness.com

Limit Order Books

These electronic platforms bring together all buyers and sellers.

- Clients can trade directly with relatively easy set-up (no need for intermediary broker). Clients can be traders directly!
- Impatient traders can trade immediately.
- Patient traders can post an order to wait for a better price.
- Traders can see the whole market (usually).
- Because it is electronic, trading is very quick (high-frequency) and algorithmic trading is easy.

All these factors tend to increase market efficiency.


Trader's Local LOBs
Order Submission Controls
Options and Customizations
Trader's Active Orders
Hotspot FX
Trader's Trading History
Active Order Details
Trade-Data Stream

How do they work?

Traders submit (upload) *orders* to the market. An order is:

- To buy or to sell ...
- a specified number of assets ...
- *either* immediately, at the best price possible
  or at a specfied price, if and when possible.

Orders to trade immediately are called *market orders*. Orders to wait until the desired price is reached (if ever!) are called *limit orders*.

Impatient traders use market orders. They want certainty of their trade price, even if they might do better by waiting. They do not want to take the risk of a worse price.

Patient traders are prepared to wait in the hope of getting a better price. They are prepared to take the risk of a worse price in return for the possibility of a better one.

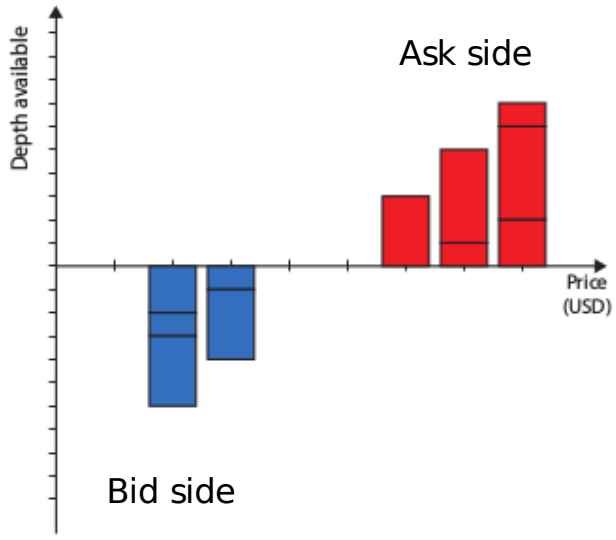
The mechanics: (1) Limit orders

First, let's see how limit orders are arranged.

The price scale is divided up into levels separated by the *tick size*. For stocks, with a typical value of \$10–\$100 the tick size might be \$0.01. For currencies, it is as small as 0.00001 units of the larger currency (eg for 1 SGD = £0.57, it would be £0.00001).

The limit orders are put into two sets of queues. There is one queue (possibly empty) at each price level.

All the buy limit orders are put on the *bid* side of the book and the sell limit orders on the *ask* side.



Q: Why are all the bid orders below all the ask orders?

The highest bid price level with a non-empty queue is called the *(best) bid price* $b(t)$ (or b_t).

The lowest ask price level with a non-empty queue is called the *(best) ask price* $a(t)$ (or a_t).

The *midprice* $m(t)$ (or m_t) is defined by

$$m(t) = \frac{b(t) + a(t)}{2}.$$

This is the price you see on TV.

The *spread* $s(t)$ (or s_t) is

$$s(t) = a(t) - b(t).$$

It is a simple measure of transaction cost in the market (the cost of a buy-sell round trip).

The *logarithmic mid-price return* between times t_1 and t_2 is

$$R_{t_1,t_2} = \log\left(m_{t_2}/m_{t_1}\right).$$

Limit order arrival and cancellation

New buy limit orders can be made anywhere below the best ask (NB) price $a(t)$. They are added to the relevant queue, usually at the back of it (so early orders are traded first).

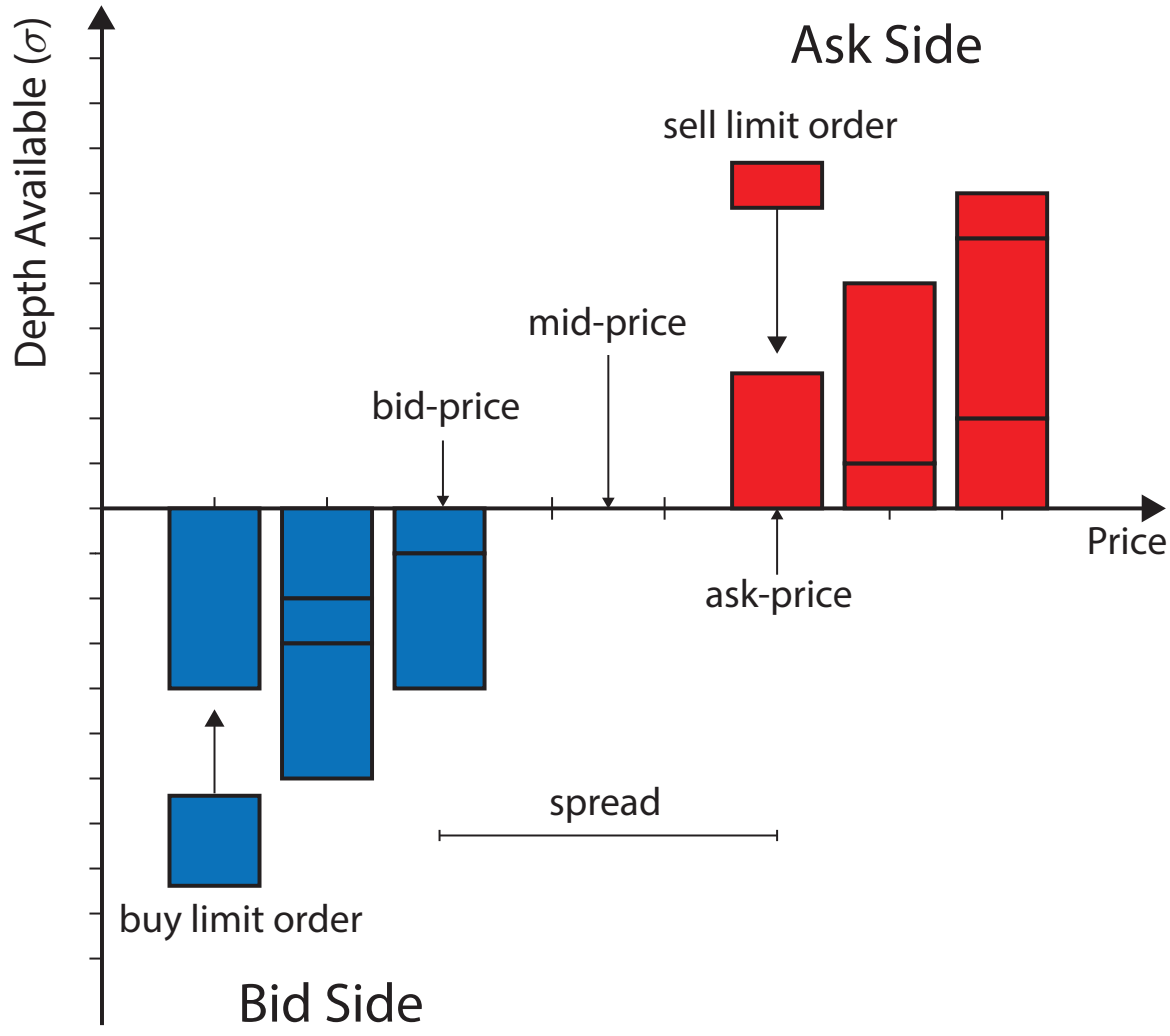
The same happens on the ask side, above the best bid price $b(t)$.

Note that this means a new bid (or ask) order can be submitted inside the spread, above $b(t)$ and below $a(t)$. That price is then the new best bid (or ask) price.

Bid and ask limit orders can be *cancelled* at any time. They are then removed from the book.

Summary of terminology

1. **tick size** smallest possible interval between consecutive prices
2. **minimum order size (lot)** smallest quantity of shares which can be traded
3. **Ask Side** all sell orders in the LOB.
4. **Ask Price** a_t the lowest price among active sell orders at time t.
5. **Bid Side** all buy orders in the LOB.
6. **Bid Price** b_t the highest price among active buy orders at time t.
7. **Mid Price** $m_t = (a_t + b_t)/2$
8. **BidAsk Spread** $s_t = a_t - b_t$
9. **Depth at a given price level p** the aggregate volume of shares to be traded, that is orders, at price level p.


Depth Available (σ)
Ask Side
sell limit order
mid-price
bid-price
Price
ask-price
spread
buy limit order
Bid Side

The mechanics: (2) Market orders

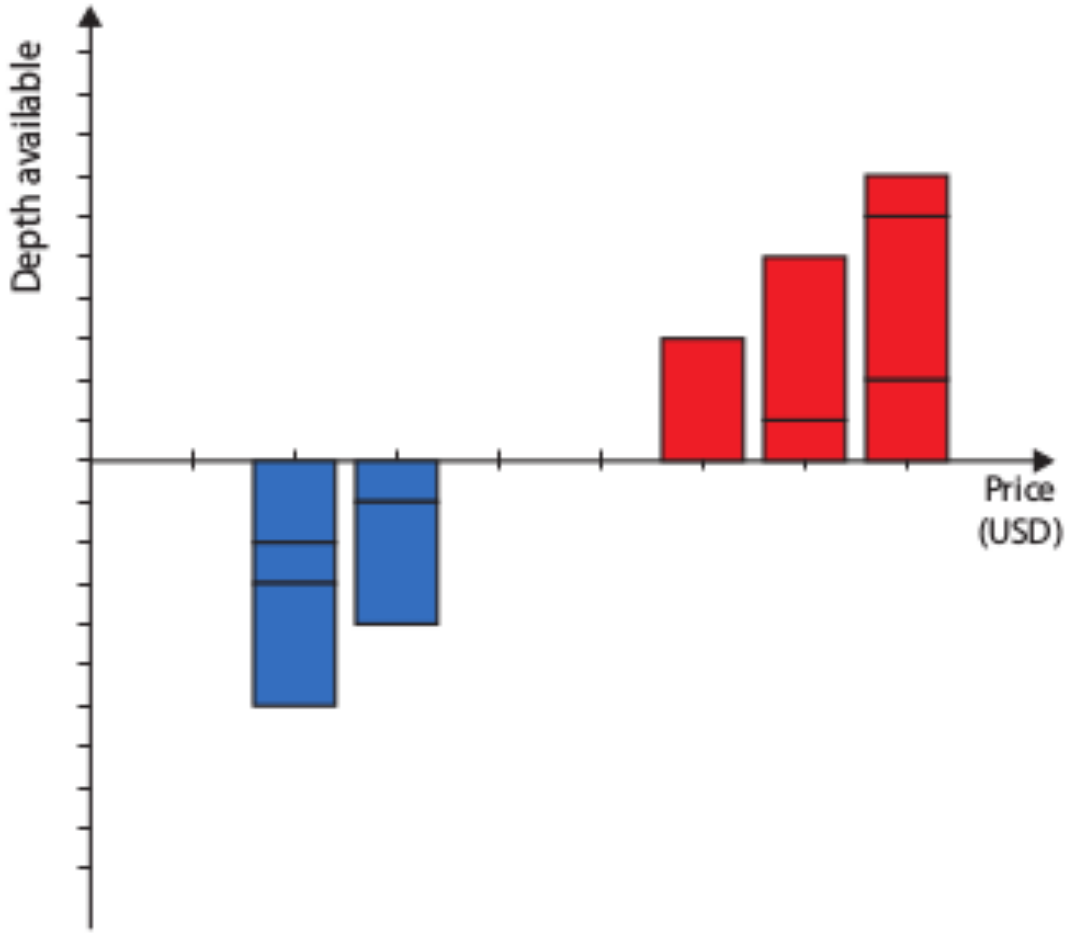
Market orders are executed immediately. Consider a market order to buy.

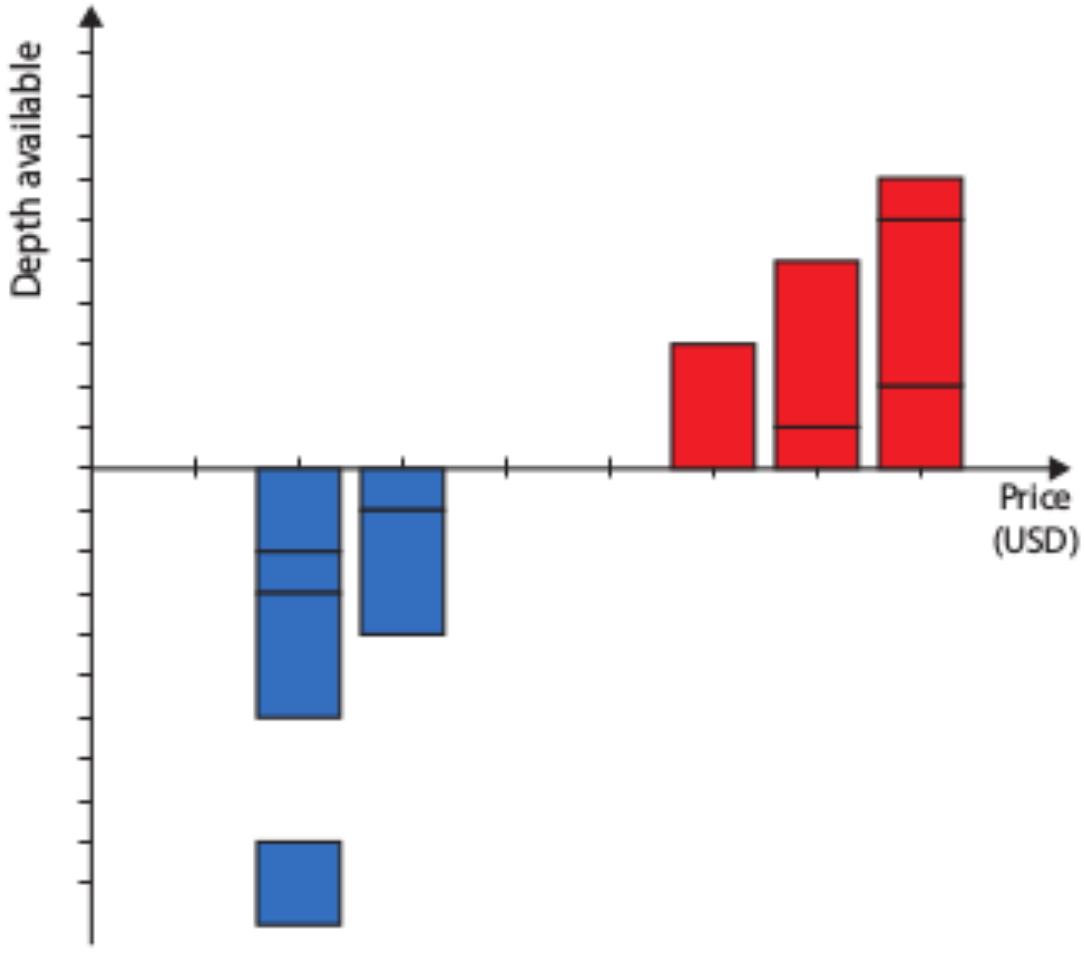
- It is matched against the queue at the best ask price $a(t)$.
- If the order size is smaller than the queue, the order is fulfilled at $a(t)$.
- If not, the remainder of the order is matched against the next non-empty queue (at a higher price) and so on.

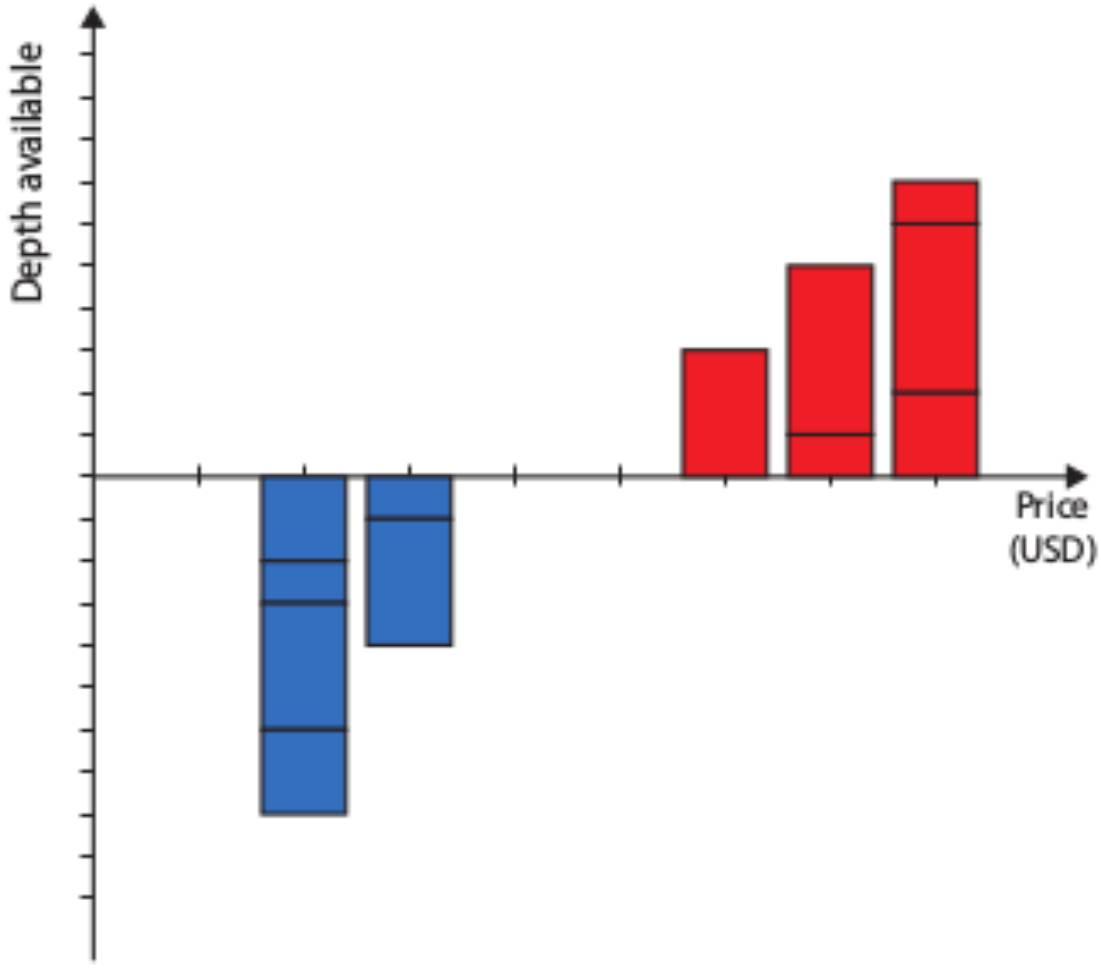
Large orders may achieve a worse price than small orders. This is called *price impact* and it has two effects:

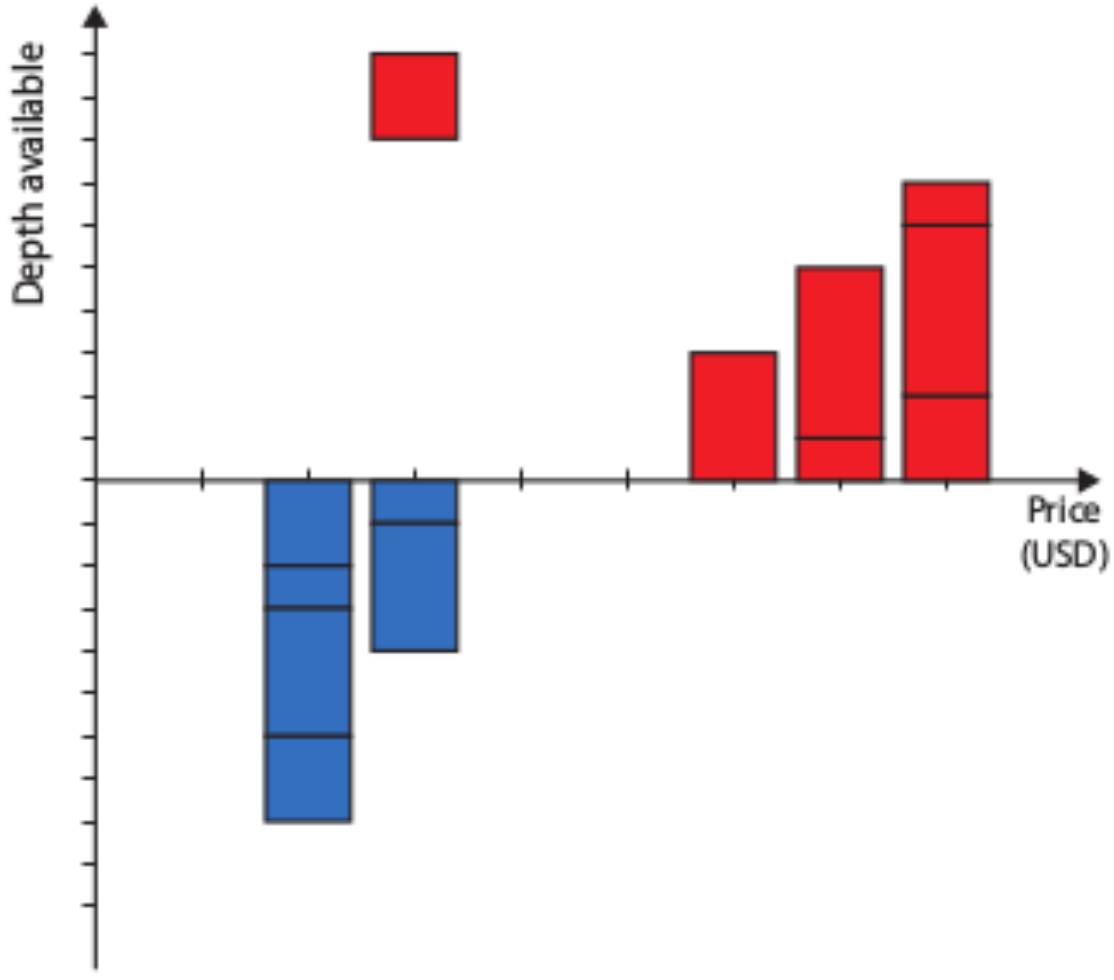
- The average price for a large order may be worse than for a small one;
- The bid/ask/mid prices may change.

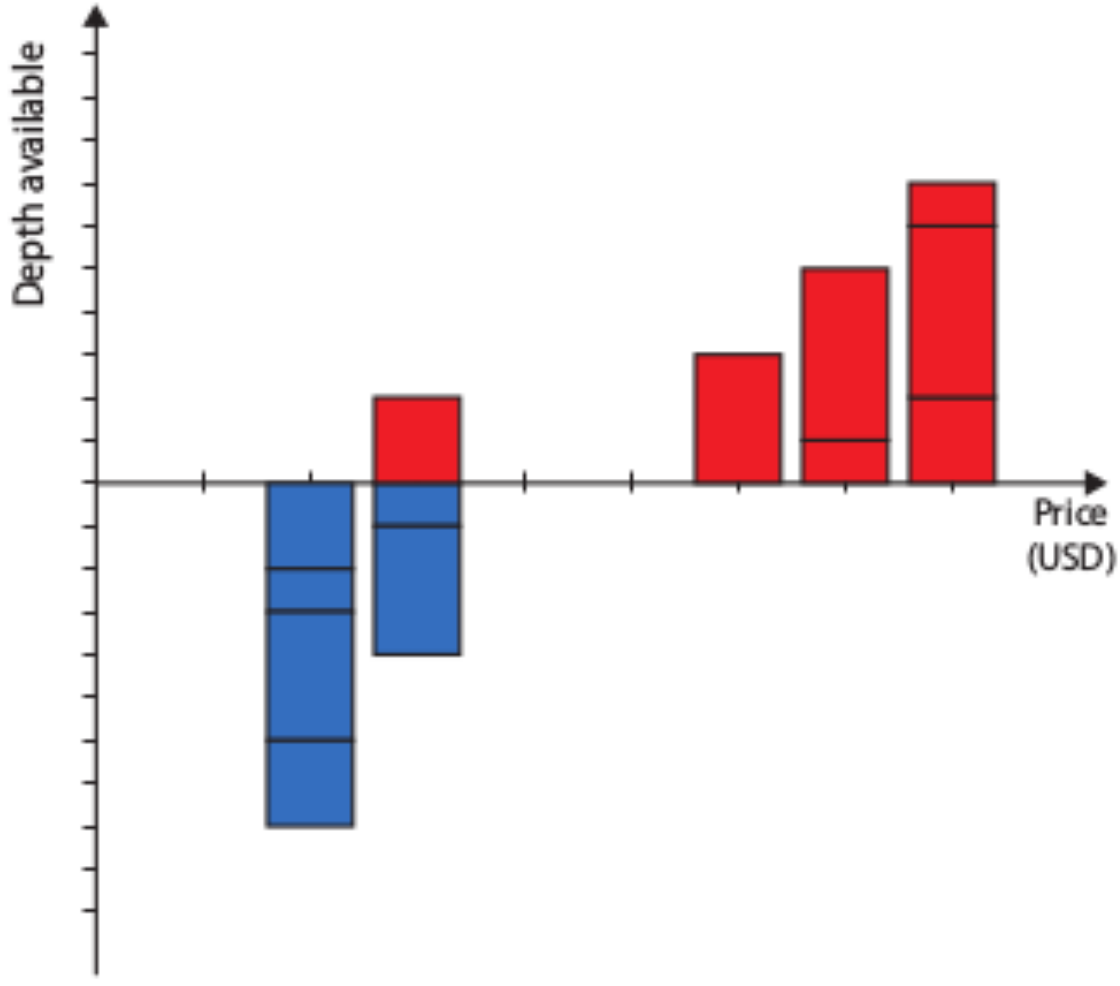
Some trades in a LOB

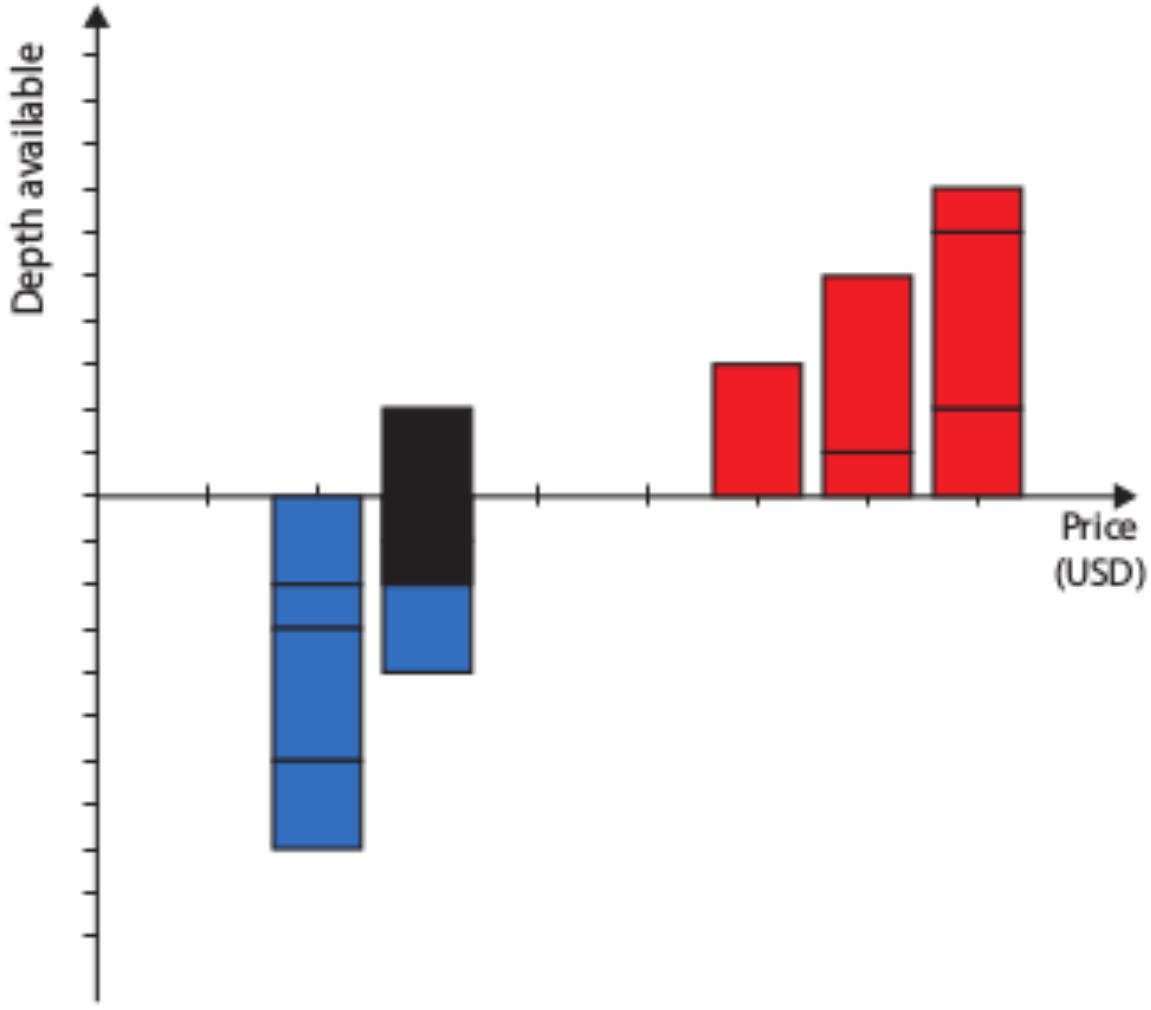


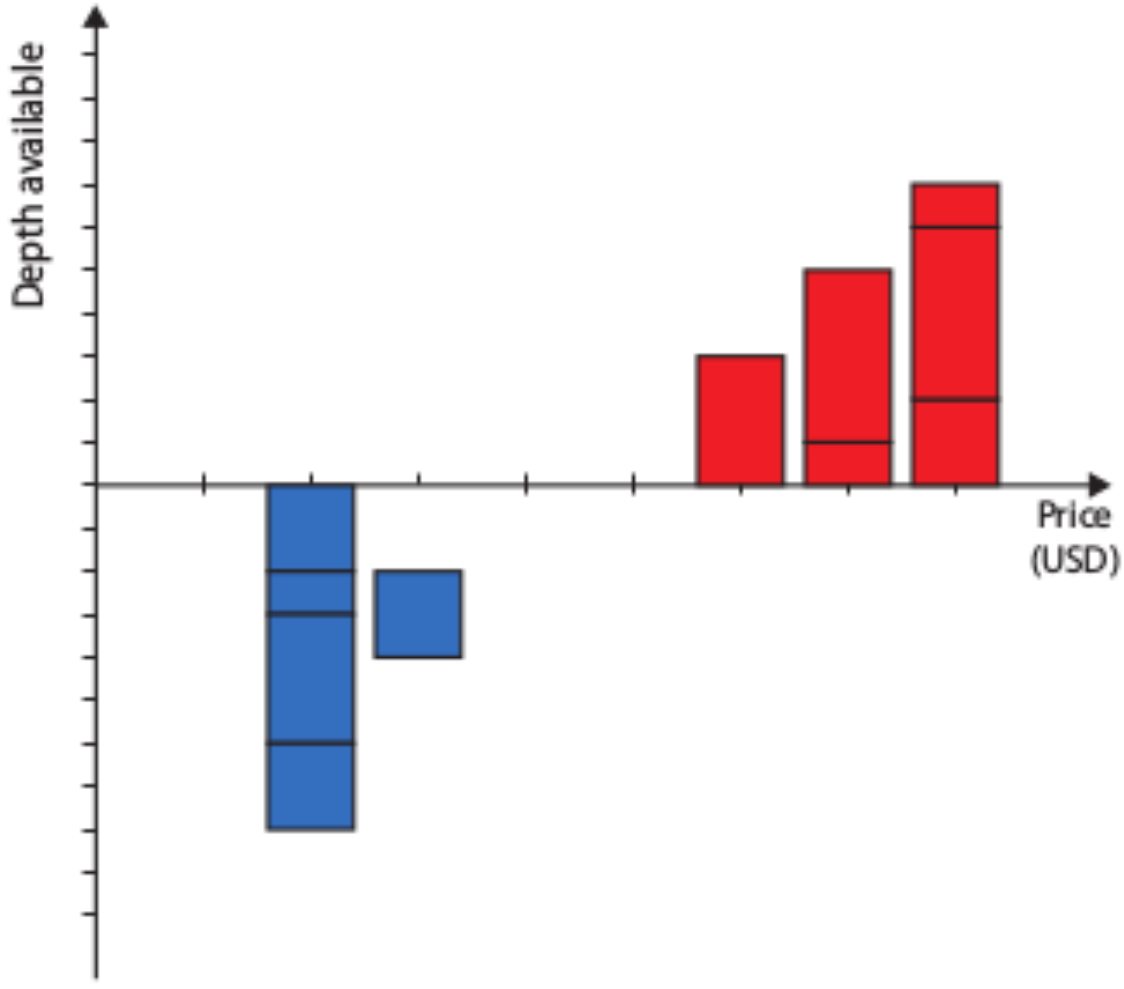

Depth available
Price
(USD)

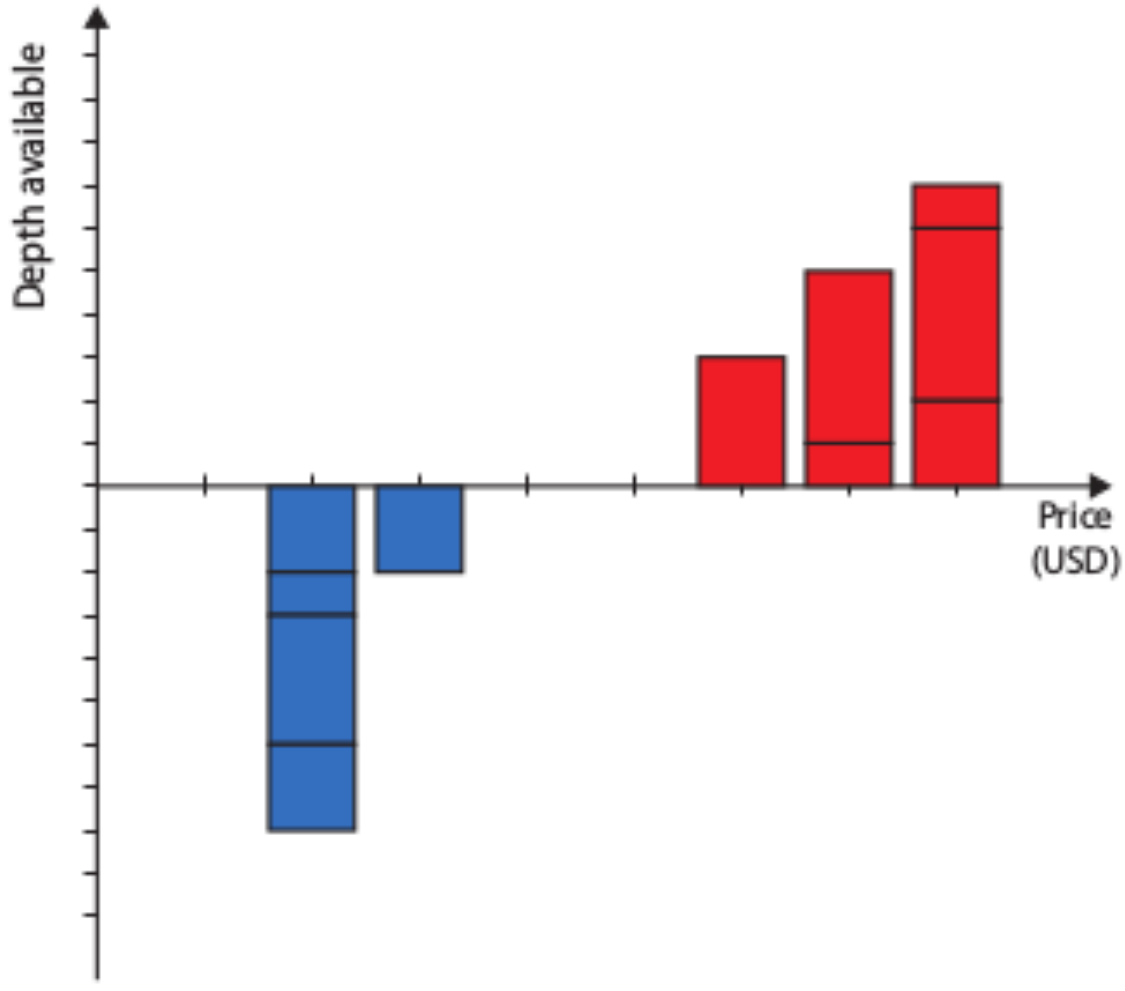

Depth available
Price
(USD)

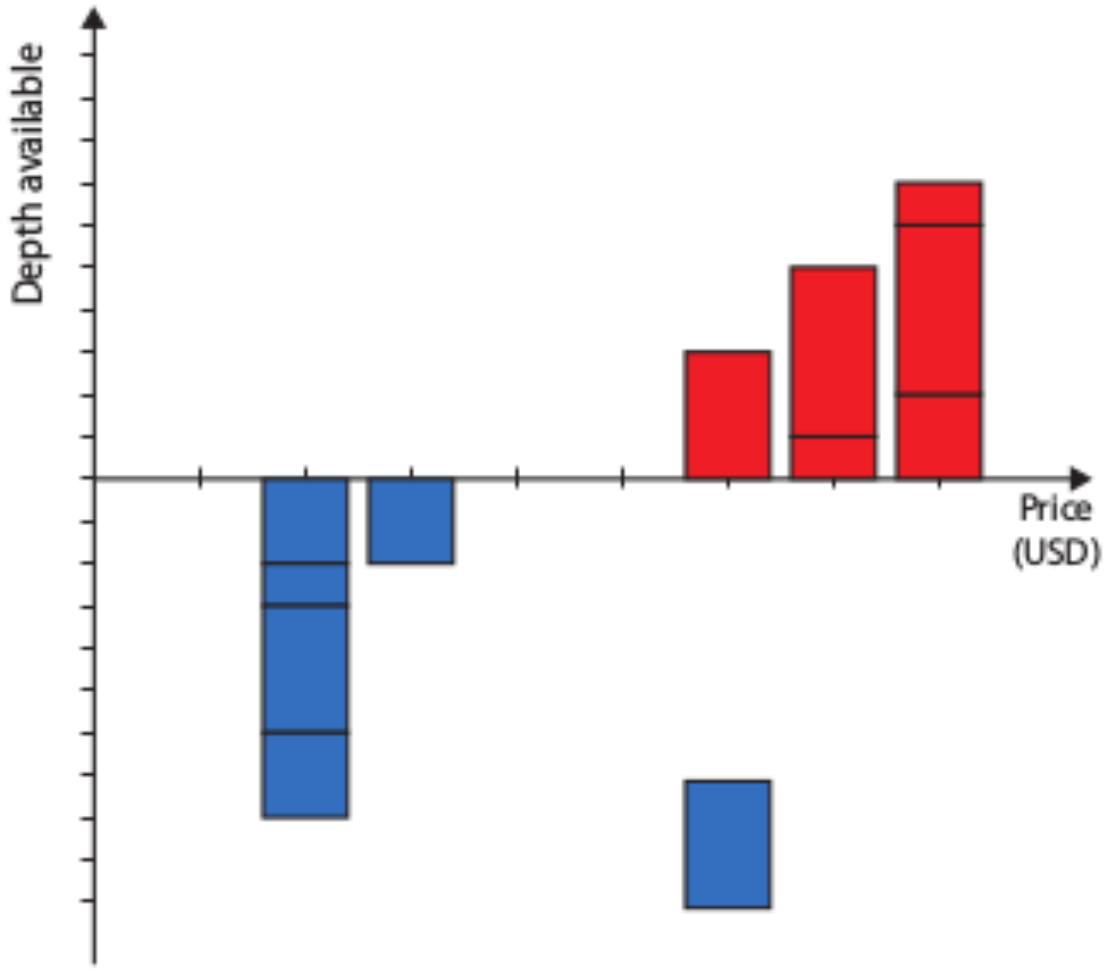

Depth available
Price
(USD)

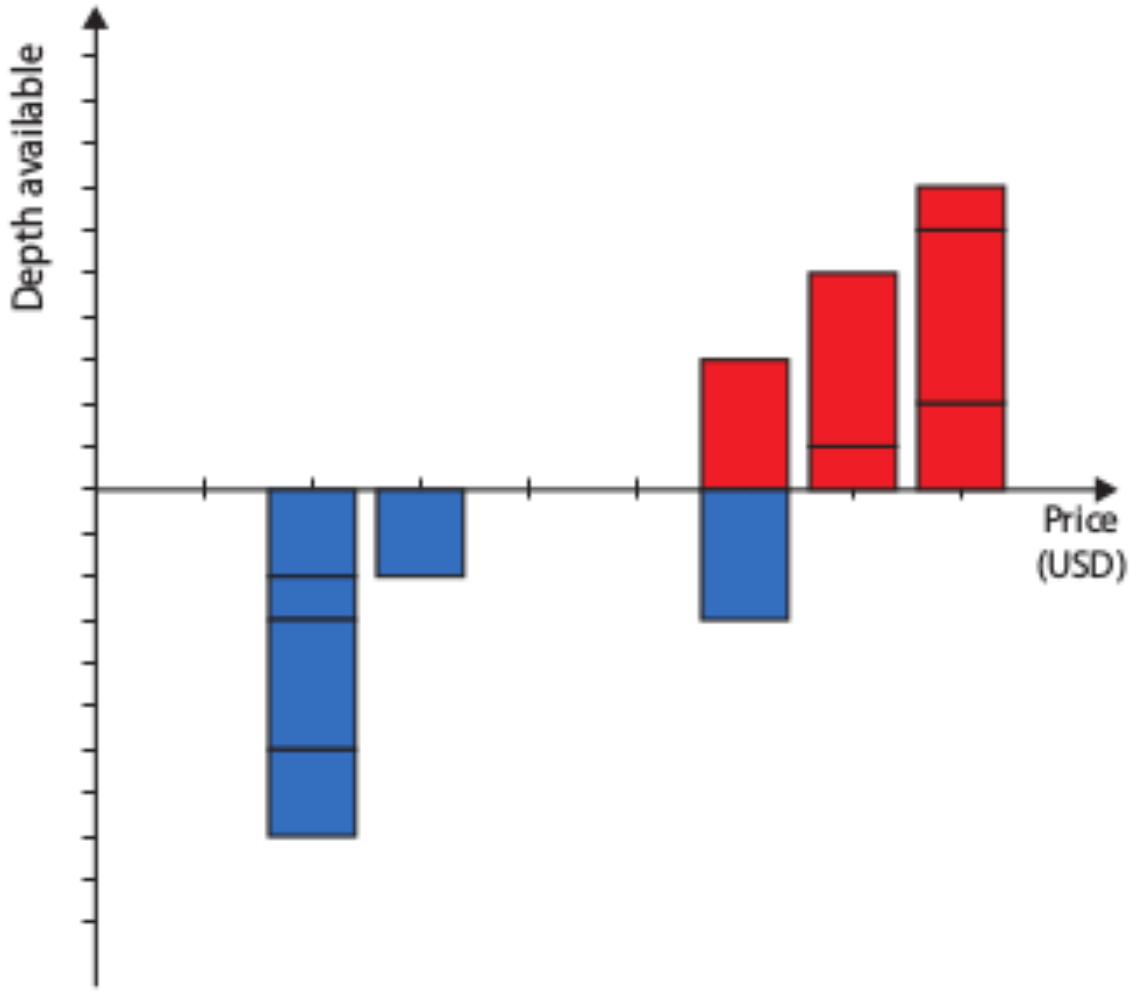

Depth available
Price
(USD)

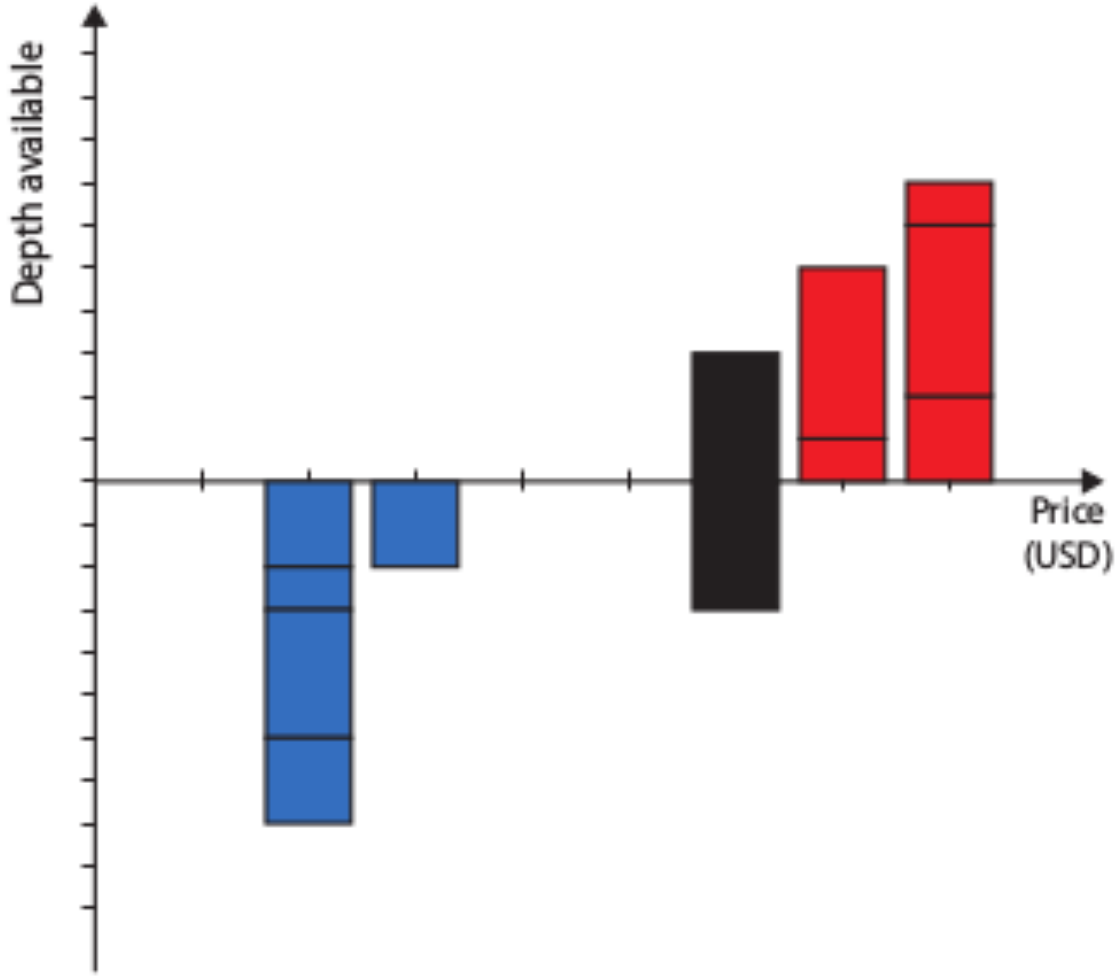

Depth available
Price
(USD)

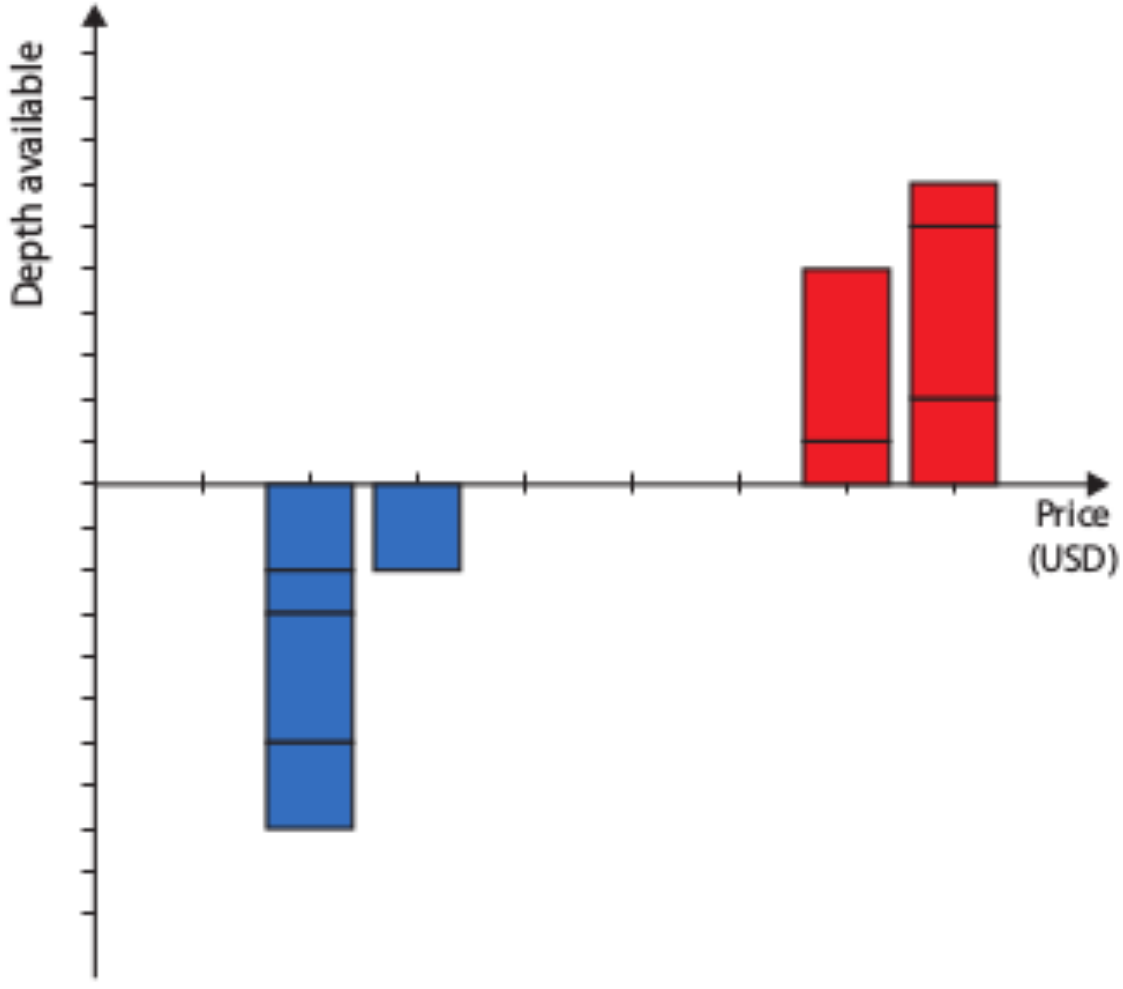

Depth available
Price
(USD)


Depth available
Price
(USD)

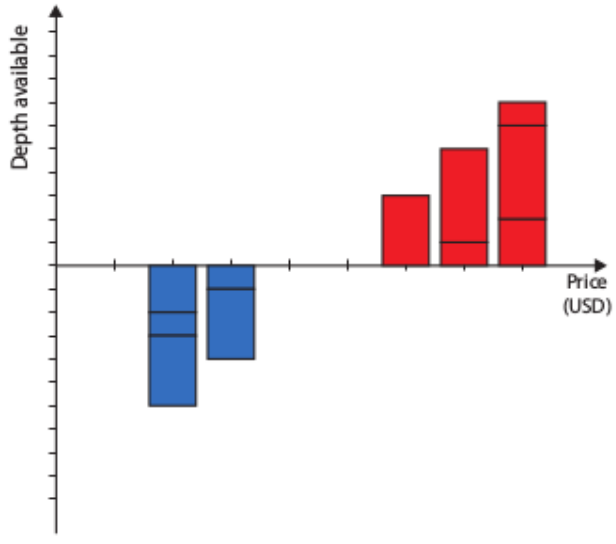

Depth available
Price
(USD)

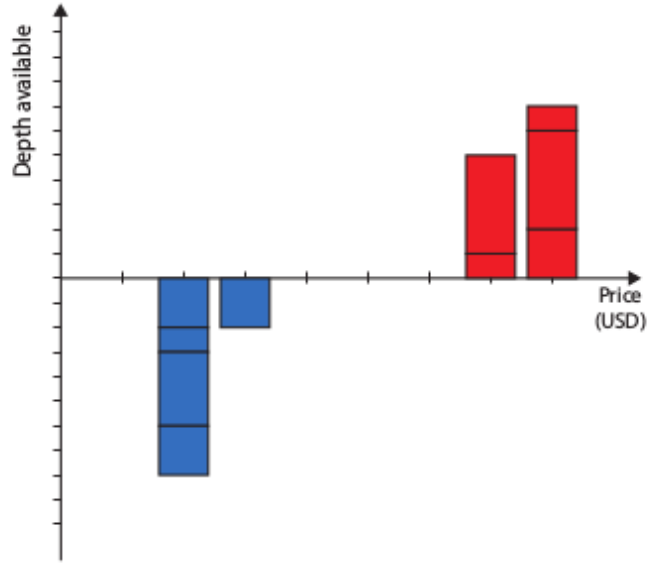

Depth available
Price
(USD)


Depth available
Price
(USD)

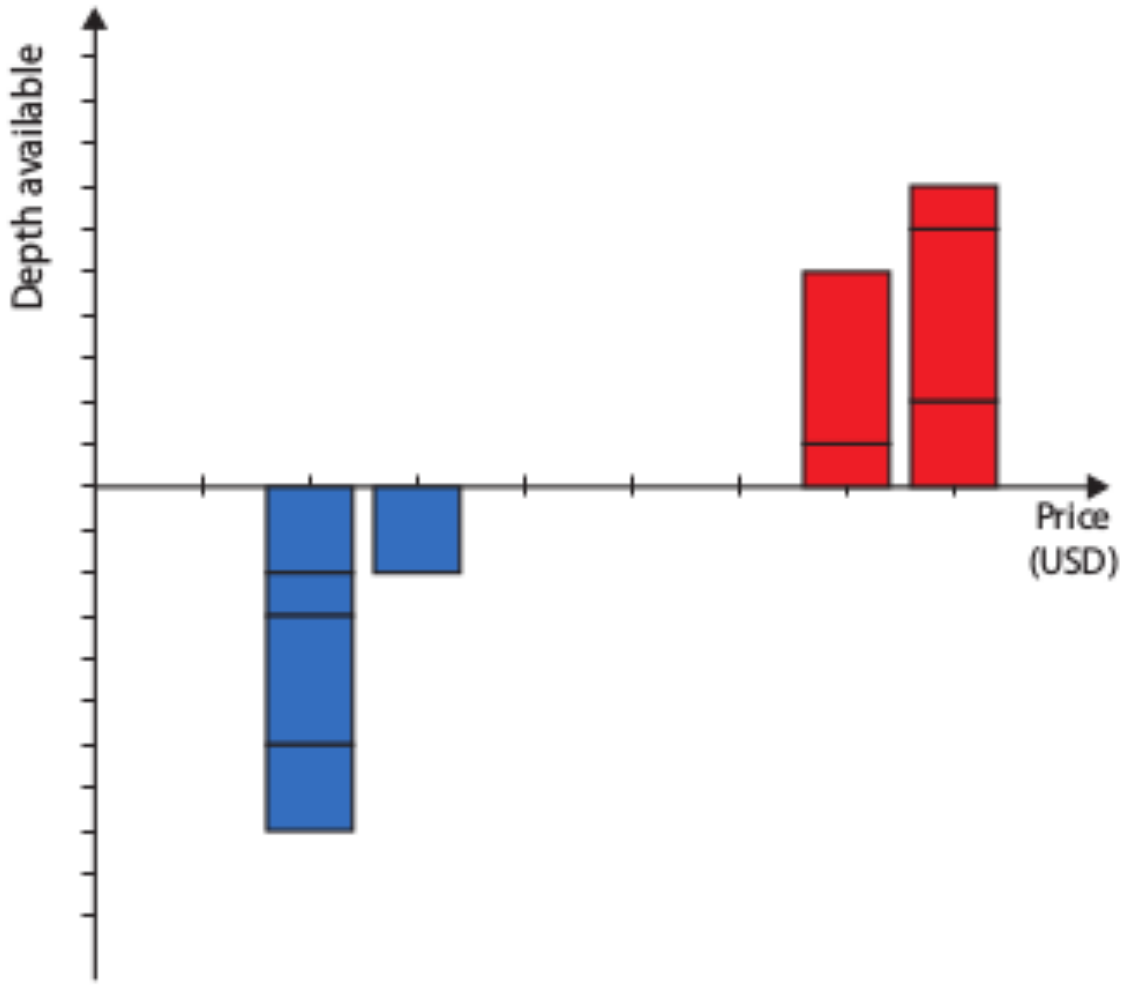

Depth available
Price
(USD)

Before and after

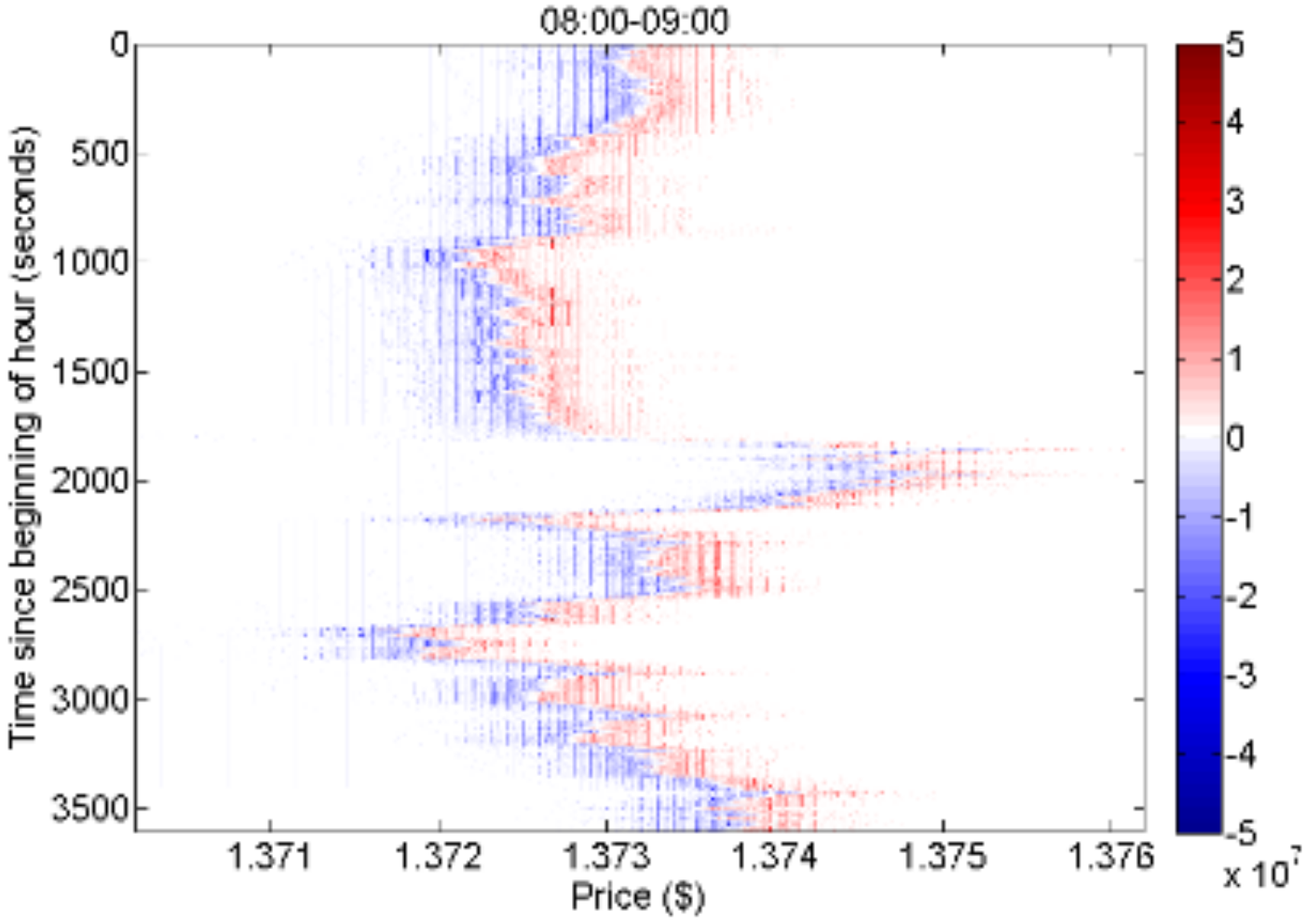




What will come next?



In real life ...


08:00-09:00
Time since beginning of hour (seconds)
0
500
1000
1500
2000
2500
3000
3500
1.371
1.372
1.373
1.374
1.375
1.376
Price ($)
5
4
3
2
1
0
-1
-2
-3
-4
-5
x 10^7

Modelling LOBs

Why model?

- *To understand price formation.* Black–Scholes and similar models are top-down 'black-box' models which say nothing about why and how prices change in response to new information, or in relation to market conditions and structure.
  For example, can we 'derive' a Black–Scholes geometric Brownian motion model by appropriately scaling a microstructure model?
- *To design trading algorithms.* High-frequency trading is made possible by LOBs. How should one trade, given knowledge of the market? What are good strategies?
- *To understand market dynamics as driven by trading strategies.* This is of interest to regulators, who want to see robust and stable markets.

Stylized facts

Any model should either reproduce or build in certain *stylized facts* (observed across many markets):

- Heavy-tailed distribution of unconditional returns: return densities are not Normal on short time scales, but tend to have power-law decay in the tail.
- Aggregational Gaussianity: over longer time scales, return densities become close to Normal.
- Fast decay of linear auto correlation of returns. Linear autocorrelations of returns are small.
- Long memory in absolute returns. This is a sort of volatility clustering.
- Long memory in order flow. That is, imbalances between buy and sell orders have long persistence.

Depth profiles and order-flow patterns

It is important to look at the 'shape' of the order book as well as the best bid and ask prices.

- The *Depth profile* is the shape of the volume of orders as a function of price. It is usually 'hump-shaped' with a maximum near the best bid/ask prices.
- The arrival rates of limit orders also vary with price.
- Almost all limit orders are cancelled, so the cancellation rate is close to the order arrival rate.

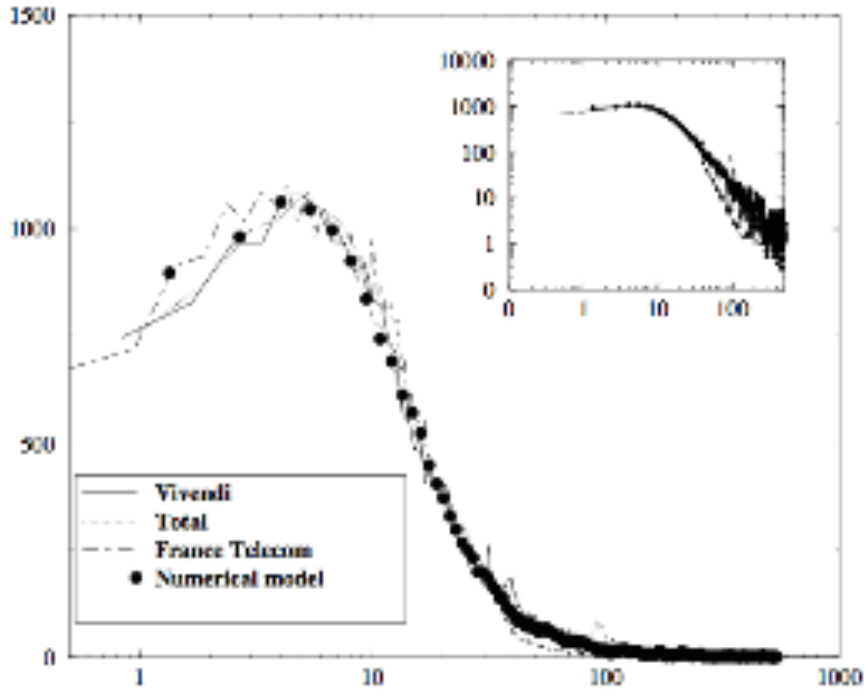
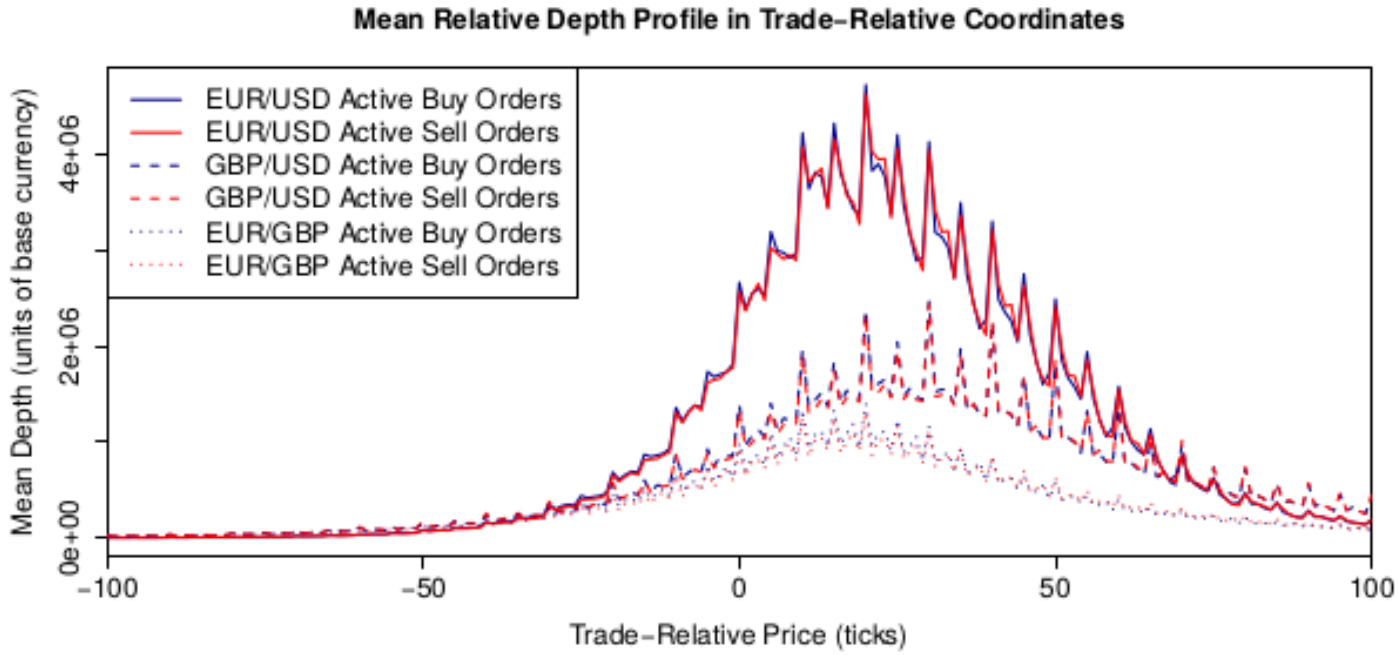
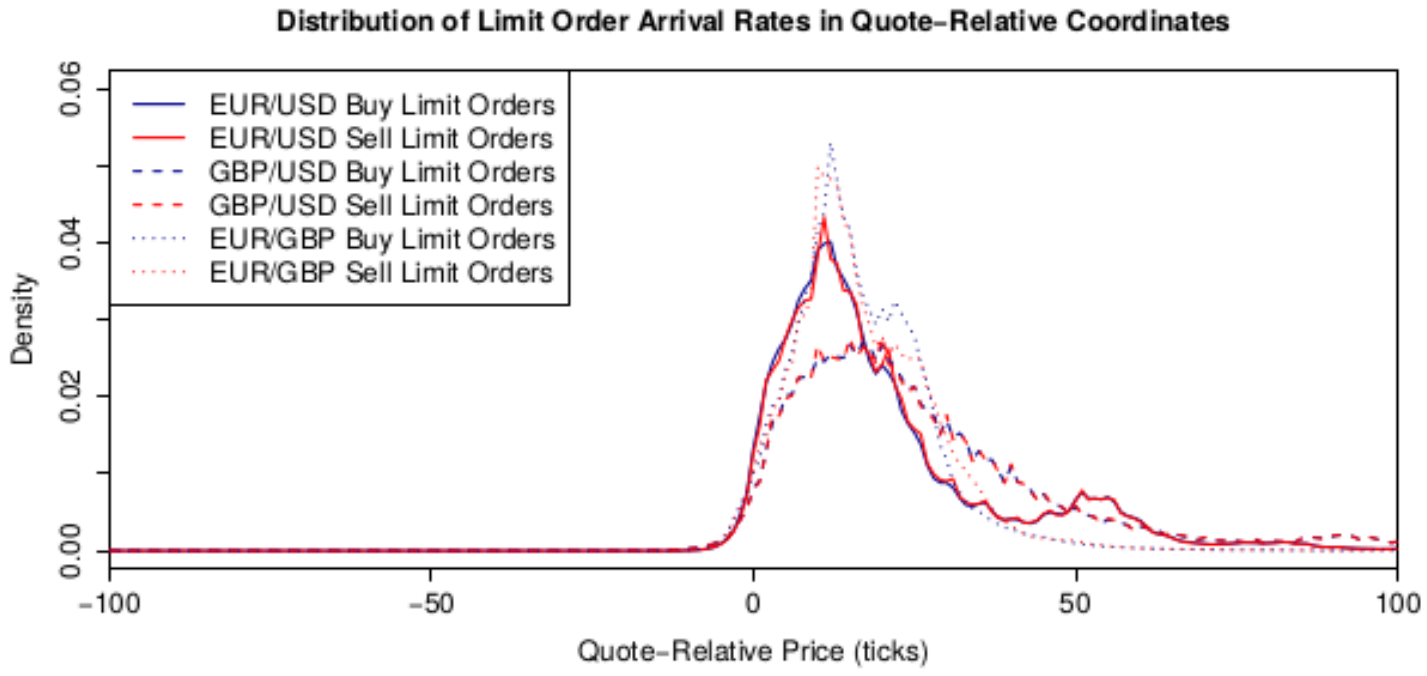


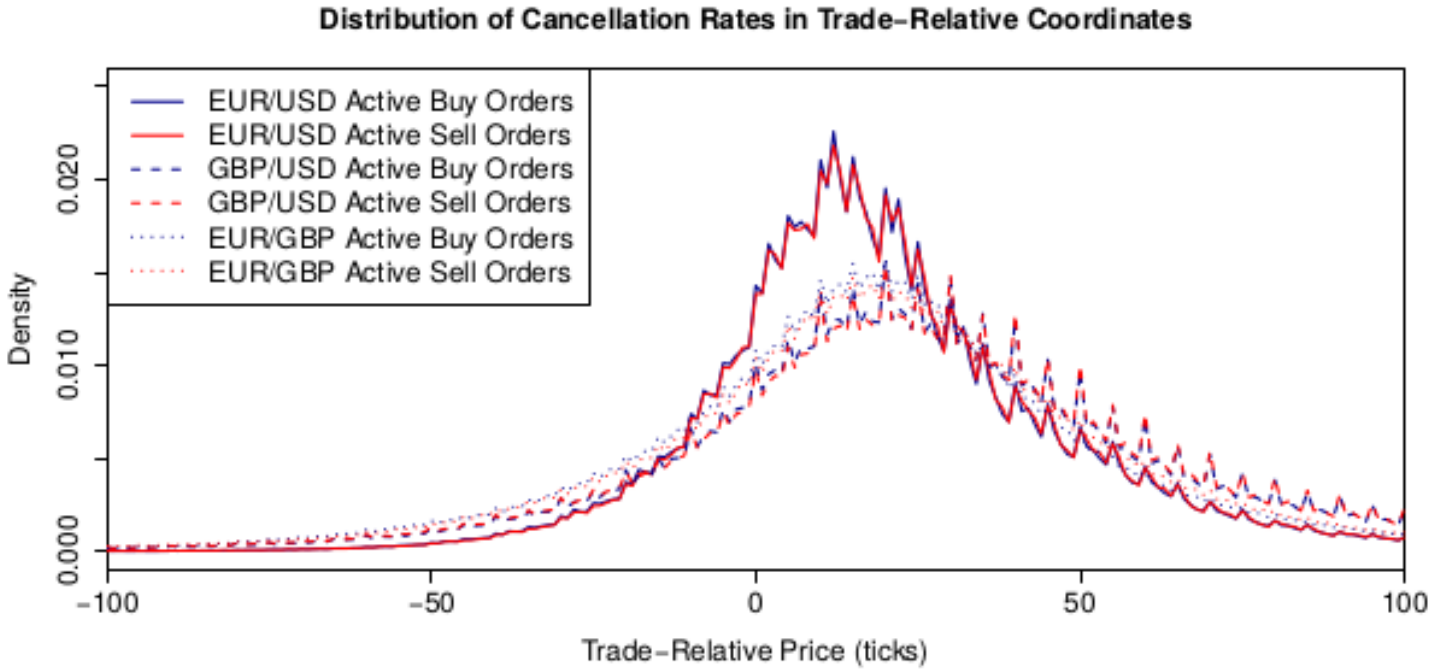
Figure: Mean relative depths profiles for Vivendi, Total, and France Telecom stocks (from "Statistical Properties of Stock Order Books: Empirical Results and Models", Bouchaud *et al.*, Quantitative Finance, 2002).



Depth in the FX market (figure: M. Gould)



Arrival rates in the FX market (M. Gould)

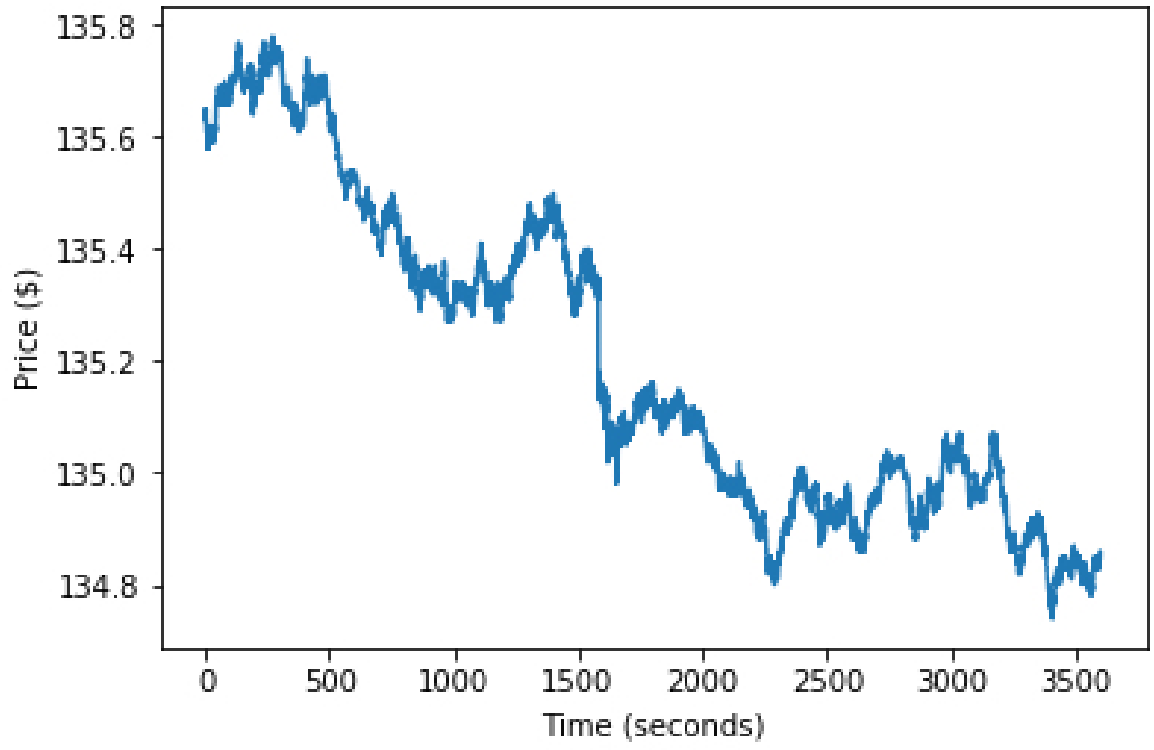


Cancellation rates in the FX market (M. Gould).

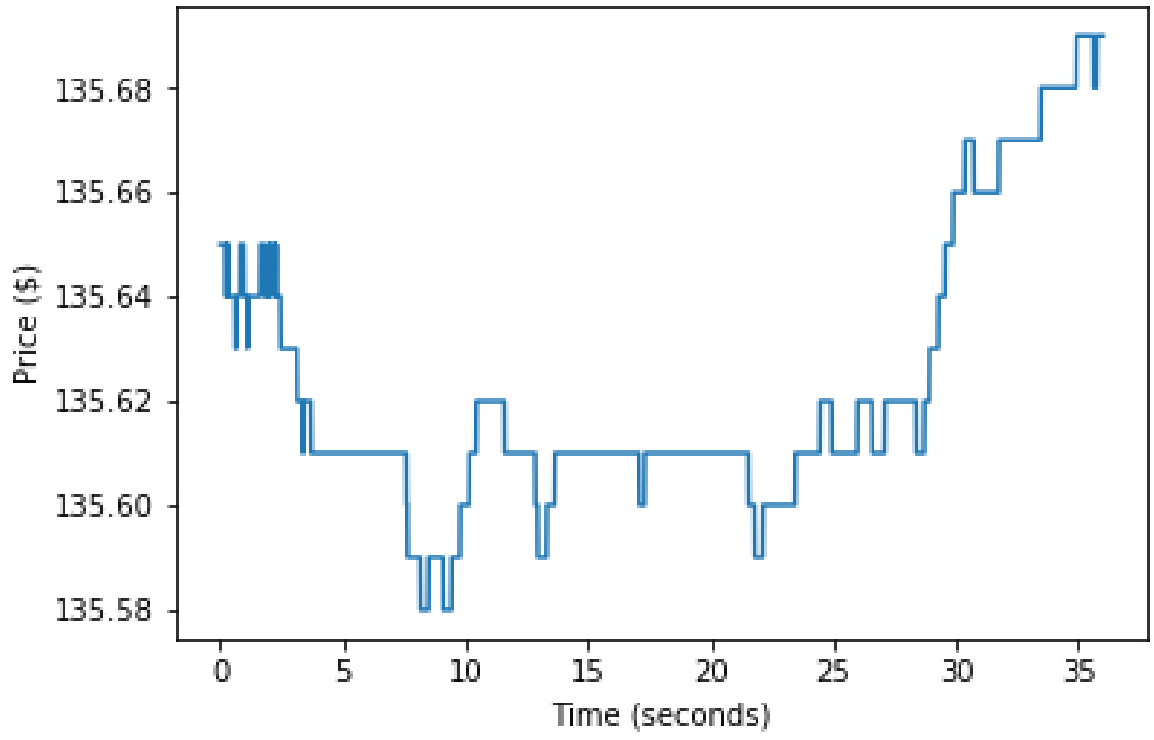
LOBSTER

Order book data can be obtained from the LOBSTER database. This has data from the NASDAQ, some of which is freely accessible.
We discuss the SPDR Trust Series I, an exchange traded fund which tracks the S& P 500, with data from June 21 2012.

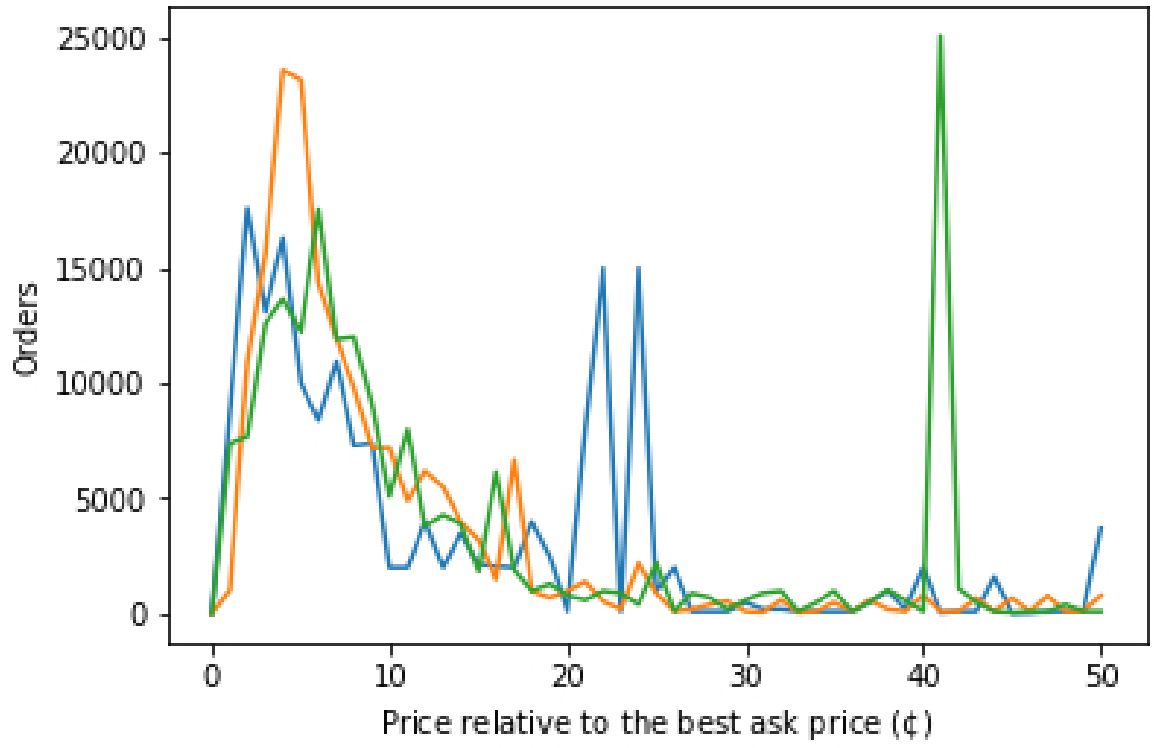
- between 9:30am and 10:30am there were 1154736 order book events and 3835 price changes.
- between 11:00am and 12:00pm there were 840549 order book events and 2453 price changes.



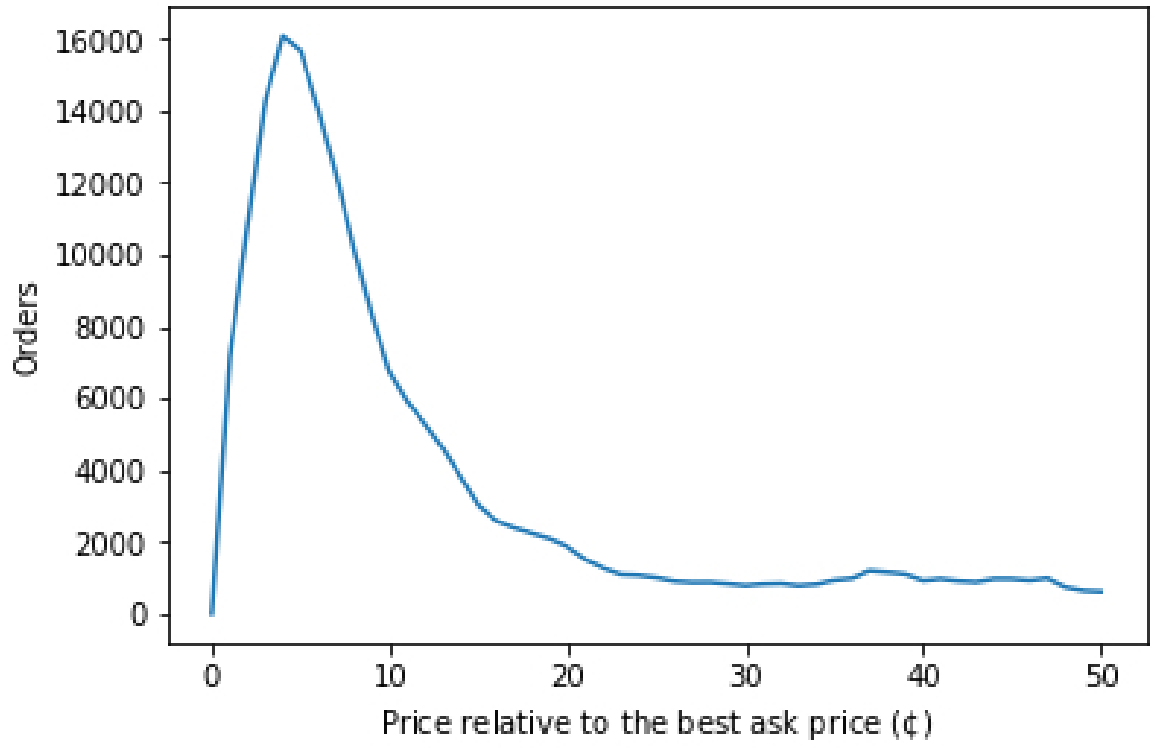
The SPDR price from 9:30-10:30.



The SPDR price for the first 36 seconds.



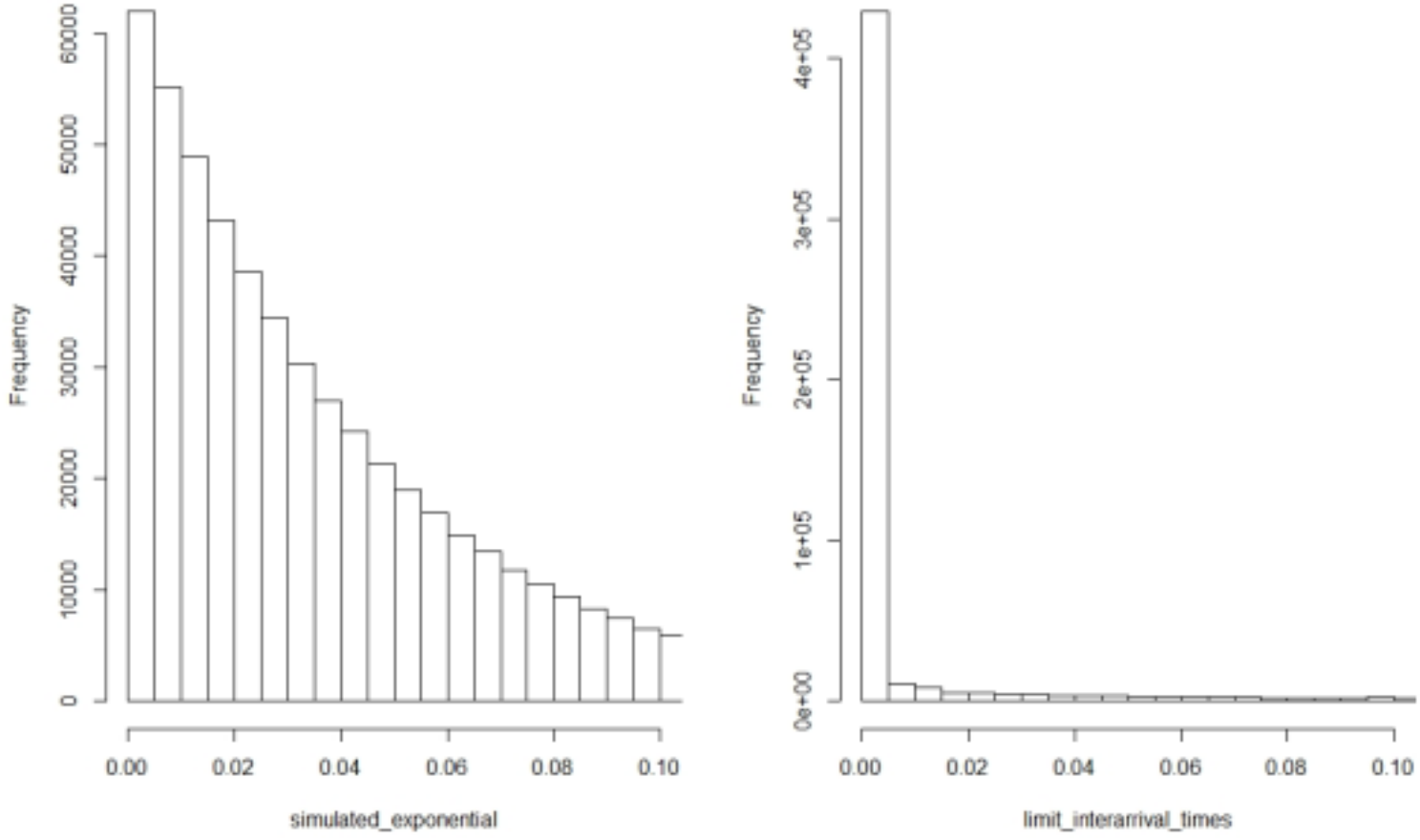
Snapshots of the profile of the order book from 9:30-10:30.



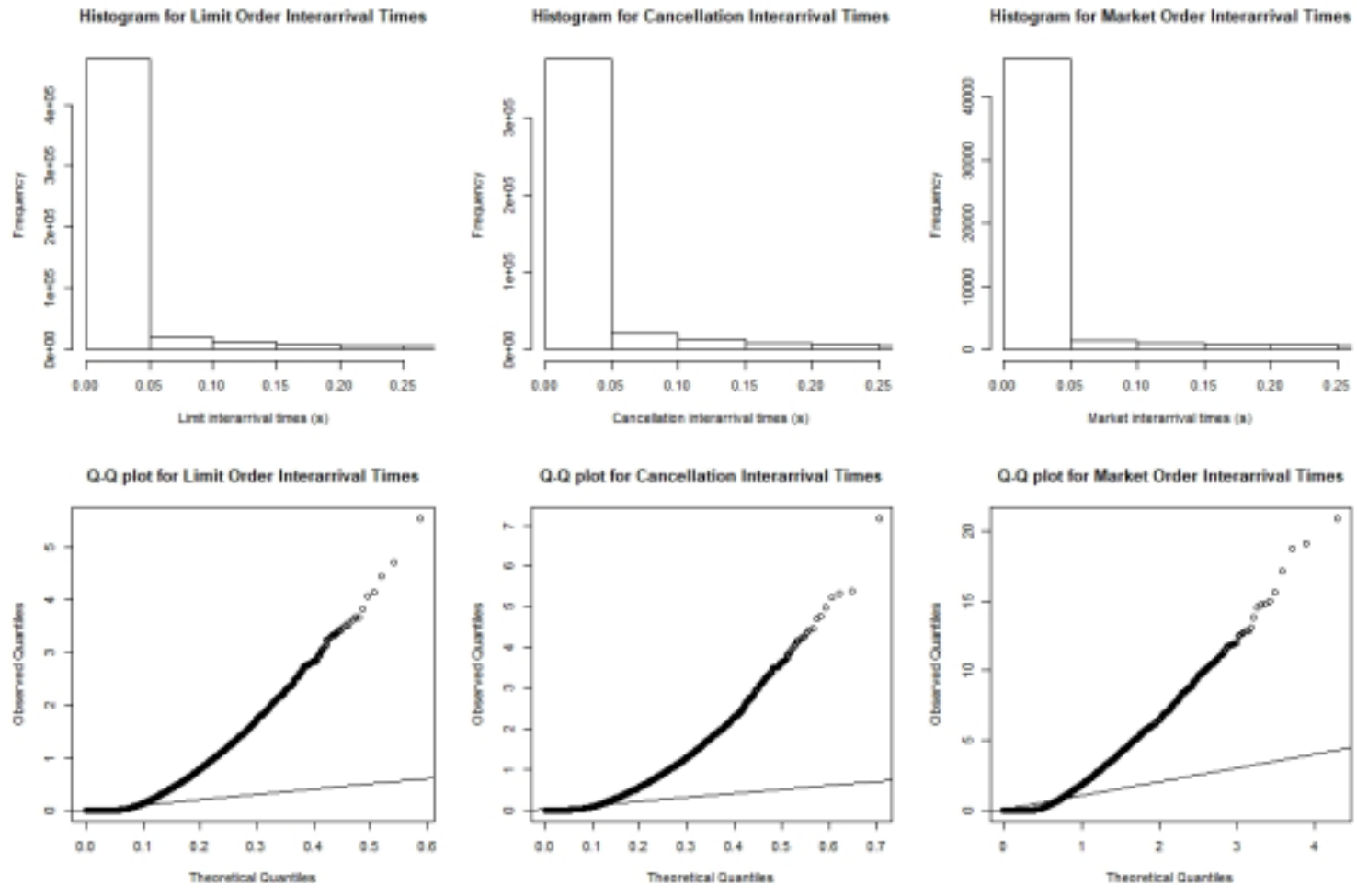
The average profile over the hour from 9:30-10:30.

Some other observations.

- For MSFT we see that order sizes are typically in lots of 100 - over 80%. Other orders are mulitples of 100. However occasionally there are very large orders of size 10000 or more.
- The time between orders is often very short with a distribution which is much steeper than an exponential.



Comparison of interarrival times for MSFT on 1/11/2018 with an exponential of the same mean (J. Mackillop)



Times between limit, cancellation and market orders (J. Mackillop)

Further issues for real LOBs

- the same shares traded on many venues smart routing is necessary
- consolidated LOB for all venues may (US) or may not (EU) be readily available
- various execution order conventions: price-time priority (FIFO), price-size priority, pro-rata priority - strategic order posting
- hidden (iceberg orders) or invisible (dark pool) liquidity fishing, price manipulation, predatory trading.

Priority

Price-time priority

- For active buy orders, priority is given to the active orders with the highest price.
- For active sell orders, priority is given to the active orders with the lowest price.
- Ties are broken by selecting the active order with the earliest submission time.

Price-time priority is an effective way to encourage traders to place limit orders. Without a priority mechanism based on time, there is no incentive for traders to show their hand by submitting limit orders earlier than is absolutely necessary.

Priority

Price-size priority

- For active buy orders, priority is given to the active orders with the highest price.
- For active sell orders, priority is given to the active orders with the lowest price.
- Ties are broken by selecting the active order with the largest size. Price-size priority is an effective way to encourage traders to place large limit orders, thereby providing liquidity to the market.

Priority

Pro-rata priority

- For active buy orders, priority is given to the active orders with the highest price.
- For active sell orders, priority is given to the active orders with the lowest price.
- When a tie occurs at a given price, each relevant active order receives a share of the matching proportional to the fraction of the depth available that it represents at that price.
- Traders in pro-rata priority LOBs are faced with the substantial difficulty of optimally selecting limit order sizes, because posting limit orders with larger sizes than the quantity that is really desired for trade becomes a viable strategy to gain priority.

Priority

Different priority mechanisms encourage traders to behave in different ways:

- Price-time priority encourages traders to submit limit orders early
- Price-size and pro-rata priority reward traders for placing large limit orders and thus for providing greater liquidity to the market.

Traders behaviour is closely related to the priority mechanism used, so LOB models need to take priority mechanisms into account when considering order flow. Furthermore, priority plays a pivotal role in models that attempt to track specific orders.

Hidden Liquidity

An *iceberg order* is a type of limit order that specifies not only a total size and price but also a visible size. Other market participants only see the visible size. Rules regarding the treatment of the hidden quantity vary greatly from one exchange to another:

- In some cases, once a quantity of at least the visible size matches to an incoming market order, another quantity equal to the visible size becomes visible, with time priority equal to that of a standard limit order placed at this time.
- Some other trading platforms, such as Currenex and Hotspot FX, allow entirely hidden limit orders. These orders are given priority behind both entirely visible active orders at their price and the visible portion of iceberg orders at their price, but they give market participants the ability to submit limit orders without revealing any information whatsoever to the market.

Dark Pools

Recently, there has also been an increase in the popularity of so-called dark pools, particularly in equities trading.

- electronic engine matching buy and sell order without routing to lit exchanges
- no information about market participants trading intentions is available to other market participants
- some dark pools are essentially LOBs in which all limit orders are entirely hidden
- other dark pools are time-priority queues of buy/sell orders (no prices specified), trading at mid-point of a reference (lit) exchange
- allows the trade of large amounts without impacting the price over 30% of all trades!

LOB Modelling: A spectrum of approaches

The modeller has to choose what to put into the model and what to try to get out of it.

- One approach is *perfect rationality* in an agent-based framework. The market is populated with agents who have certain trading goals and act to maximise utility.
  - Complex with many unobservable parameters;
  - Not easy to construct realistic strategies;
  - Tend to come from the economics literature. See Gould et al. Quantitative Finance **13**, 1709–1742, 2013 for a review.

Zero-intelligence models

- *Zero-intelligence* models are probabilistic models for order flow based on observed statistical properties.
  - Specify processes to depend on the state of the book.
  - Much easier to calibrate ...
  - But is it realistic to ignore agents' intentions and strategies?

See models by Smith et al (Quantitative Finance **3**, 481–514, 2003) and Cont et al. (Operations Research **58**, 549–563, 2010).

Zero-intelligence (continued)

Zero-intelligence models are essentially a collection of queues representing the set of orders at each price tick. Order arrival is governed by a Poisson process for each queue:

- Buy and sell orders arrive at a rate which, for simplicity, may be constant;
- Limit orders and cancellations arrive at rates that depend on position in the book (to mimic the stylized facts above).

The state space is extremely large and the dynamics can be very complicated. Nevertheless some good outputs can be obtained. It is possible to incorporate autocorrelation of the order flows using Hawkes processes.

In particular, over long timescales, in some situations the mid-price evolution tends to a Brownian Motion.

However, it is not easy to study trading strategies in them.

Hybrid approaches

Hybrid approaches usually have one trader (you!) operating in a noisy environment. Their main use is to guide optimal trading: how best to sell a large order given the price impact of your trade.

A typical set-up has:

- A Brownian Motion representation of the mid-price;
- A simple (summary) representation of the bid and ask sides of the book, for example
  - Simple stochastic processes representing the queues at the best bid and ask (proxies for the whole order profile);
  - Parametrized representations of the price impact of trading: how does your trading affect the price over and above its intrinsic noise?
- A notion of optimality for the trader: maximise a measure of return while achieving the trading goal.

In the next lectures you'll hear a lot more about these models!

Zero-intelligence models

We consider this now in more detail. As the LOB is considered as a complex queueing system we recall some queueing thoery ideas and how they need to be extended to handle the issues that arise from LOBs.

- We recall Poisson processes and how they can be used in simple queueing models
- We then think about how this may fit with the overall modelling objective.
- We consider some approximations via continuous stochastic processes

Poisson processes

We review the basics of *Poisson processes* (PPs).

In one dimension, these are the most fundamental continuous-time processes with a finite number of jumps.

How do you distribute a countable set of points on $\mathbb{R}$ so that every interval of length t has, on average, λt points (here $\lambda > 0$ is a constant)?

As usual, we have a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ in the background.

PPs count events

Consider events (example: radioactive decay) such that

1. They can occur independently again and again;
2. An event is equally likely to occur in any small time interval.

Let N_t be the number of events that occur by time t, with $N_0 = 0$. We assume $N_{t+h} - N_t$ is independent of N_t for all $t, h > 0$.
Take any small time interval $[t, t+h)$. Assume that for every n we have

$$\mathbb{P}[N_{t+h} = n+1 | N_t = n] = \lambda h + o(h)$$

as $h \to 0$, where the constant λ is called the *intensity*. Then

$$\mathbb{P}[N_{t+h} = n | N_t = n] = 1 - \lambda h + o(h).$$

We call N_t a Poisson process with intensity λ.

Basic properties

Clearly:

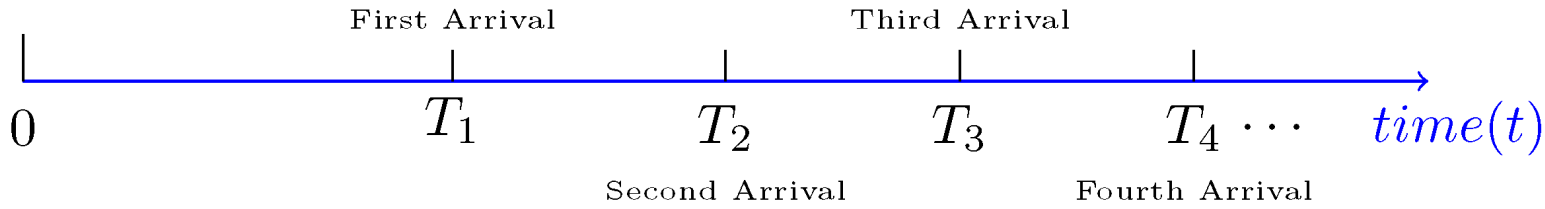
- N_t takes values in $\{0, 1, 2, 3, \ldots\}$, with $N_0 = 0$.
- N_t is constant except at times when an event occurs. We call these the *jump times*. They are random variables. Write T_i for the time of the ith occurrence, with $T_0 = 0$. So N_t has a jump of $+1$ at time T_i:

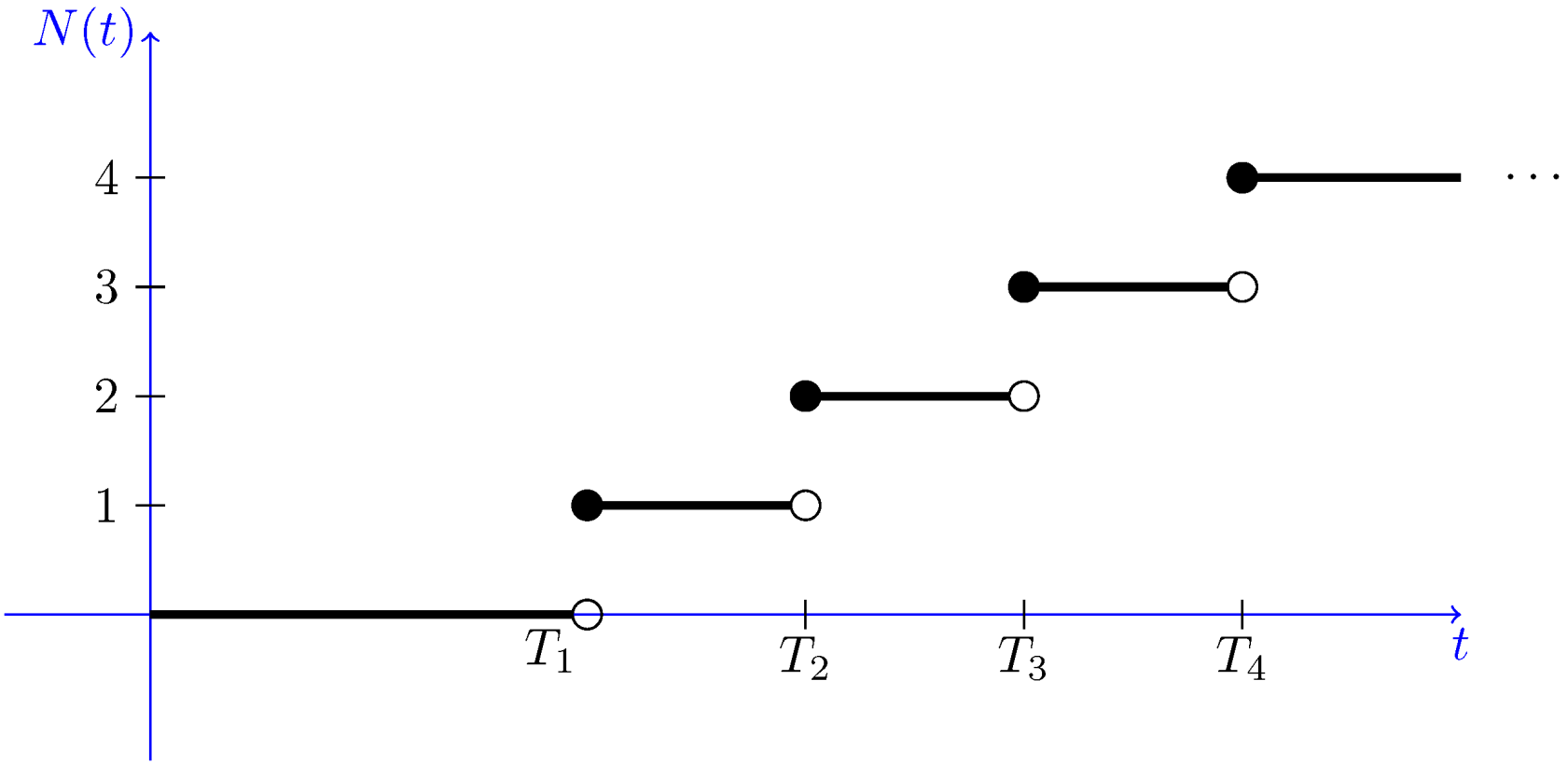
$$N_{T_i^+} = N_{T_i^-} + 1.$$

(here $N_{T_i^-}$ means $\lim_{t \uparrow T_i} N_t$ and so on). So the process N_t is cadlag (right-continuous with left limits).

We write $\mathrm{d}N_t$ for the increment of N_t and then, for each t,

$$\mathrm{d}N_t = \begin{cases} 1 & \text{with probability } \lambda\,\mathrm{d}t, \\ 0 & \text{with probability } 1 - \lambda\,\mathrm{d}t. \end{cases}$$


First Arrival
Third Arrival
0
T_1
T_2
T_3
$T_4 \cdots$
$time(t)$
Second Arrival
Fourth Arrival


$N(t)$
4
3
2
1
$\cdots$
T_1
T_2
T_3
T_4
t

A heuristic calculation for the expectation of N_t is as follows. Let

$$e_t = \mathbb{E}[N_t | N_0 = 0],$$

so that $e_0 = 0$. Then,

$$\begin{aligned} \mathrm{d}e_t &= \mathrm{d}\mathbb{E}[N_t] \\ &= \mathbb{E}[\mathrm{d}N_t] \\ &= \lambda \, \mathrm{d}t, \end{aligned}$$

so we have

$$e_t = \lambda t.$$

This is consistent with the intuition that the events (jumps) are 'uniformly distributed in time'.

Distribution of the number of jumps

We now calculate the PDF of the number of jumps in an interval of length t. That is, we want to find

$$p_n(t) = \mathbb{P}[N_t = n], \qquad n = 0, 1, 2, \ldots.$$

Use the probability generating function

$$\begin{aligned} G_t(s) &= \mathbb{E}[s^{N_t}] \\ &= \sum_{n=0}^{\infty} p_n(t) s^n. \end{aligned}$$

Note that (i) $N_{t+h} = N_t + N_{t+h} - N_t$, (ii) $N_{t+h} - N_t$ is independent of N_t and 'has the same distribution as' N_h.
Now we calculate

$$\begin{aligned} G_{t+h}(s) - G_t(s) &= \mathbb{E}[s^{N_{t+h}}] - G_t(s) \\ &= G_t(s)\mathbb{E}[s^{N_h}] - G_t(s) \qquad \text{((i) \& (ii) above)} \\ &= G_t(s)((1-\lambda h)\times 1 + \lambda h \times s) - G_t(s) + o(h) \\ &= \lambda(s-1)G_t(s)h + o(h). \end{aligned}$$

divide by h and let $h \to 0$ gives the ODE

$$\frac{dG_t(s)}{dt} = \lambda(s-1)G_t(s).$$

As $G_t(1) = 1$ we solve to find

$$\begin{aligned} G_t(s) &= \mathrm{e}^{\lambda t(s-1)} \\ &= \sum_{n=0}^{\infty} \left(\mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} \right) s^n. \end{aligned}$$

The number of jumps is Poisson distributed with parameter λt:

$$N_t \sim \mathrm{Po}(\lambda t).$$

Q: What is $\mathrm{var}[N_t]$?

Q: show that the sum of independent Poisson random variables with parameters λ and μ is Poisson with parameter $\lambda + \mu$.

Distribution of the waiting times

The *waiting times* $\tau_i = T_{i+1} - T_i$ between jumps are IID and have the same distribution as τ_1.
We can see that

$$\mathbb{P}(\tau_1 > t) = \mathbb{P}(N_t = 0) = e^{-\lambda t}.$$

Thus τ_1 has the exponential distribution

The waiting times are exponentially distributed: $\tau_i \sim \mathrm{Exp}(\lambda)$.

Q: Show that the waiting times have the 'lack of memory property':

$$\mathbb{P}[\tau_i > s + t | \tau_i > s] = \mathbb{P}[\tau_i > t].$$

'Itô's' formula

What is the evolution of a function $f(N_t)$? If we write $f_t = f(N_t)$, then clearly f_t is constant between jumps, because N_t is.
If a jump occurs at time t,

$$\begin{aligned} \mathrm{d}f_t &= f_{t^+} - f_{t^-} \\ &= f(N_{t^+}) - f(N_{t^-}) \\ &= f(N_{t^-} + 1) - f(N_{t^-}). \end{aligned}$$

Note that everything is evaluated at time t^-: non-anticipating!
Remembering that $\mathrm{d}N_t = 0$ except at jumps, we can write

$$\mathrm{d}f_t = (f(N_t + 1) - f(N_t))\,\mathrm{d}N_t$$

because the bracketed term is only evaluated when $\mathrm{d}N_t \neq 0$, ie at a jump.

Compensated processes

We use martingales a lot in finance. However a standard Poisson process is not a martingale because, for $t > s \geq 0$,

$$\mathbb{E}[N_t | N_s] = N_s + \lambda(t - s) > N_s.$$

The solution is to subtract off the expectation of N_t: the process

$$M_t = N_t - \lambda t$$

in indeed a martingale. The term $-\lambda t$ is called the *compensator* of N_t and M_t is called a *compensated Poisson process*.

Q: Let $f_t = (N_t)^2$. By Ito what is $\mathrm{d}f_t$? Use this and the compensator to calculate $\mathrm{var}[N_t]$.

Building more complex processes

We can make many models using PPs in various ways:

- Add a deterministic function of time
- Add other processes; for example

$$X_t = \alpha N_t + \sigma W_t$$

where W_t is a Brownian Motion and α, σ are constants. This is an example of *jump-diffusion*.

- Let the jumps be different from 1. Eg they could be independent samples from some distribution. So we could have a process X_t satisfying

$$\mathrm{d}X_t = J_t \mathrm{d}N_t$$

and every time an event occurs ($\mathrm{d}N_t = 1$), X_t changes by J_t. This is a *compound Poisson process*.
Q: If the Js are IID with mean μ_J, what is the compensator of X_t?

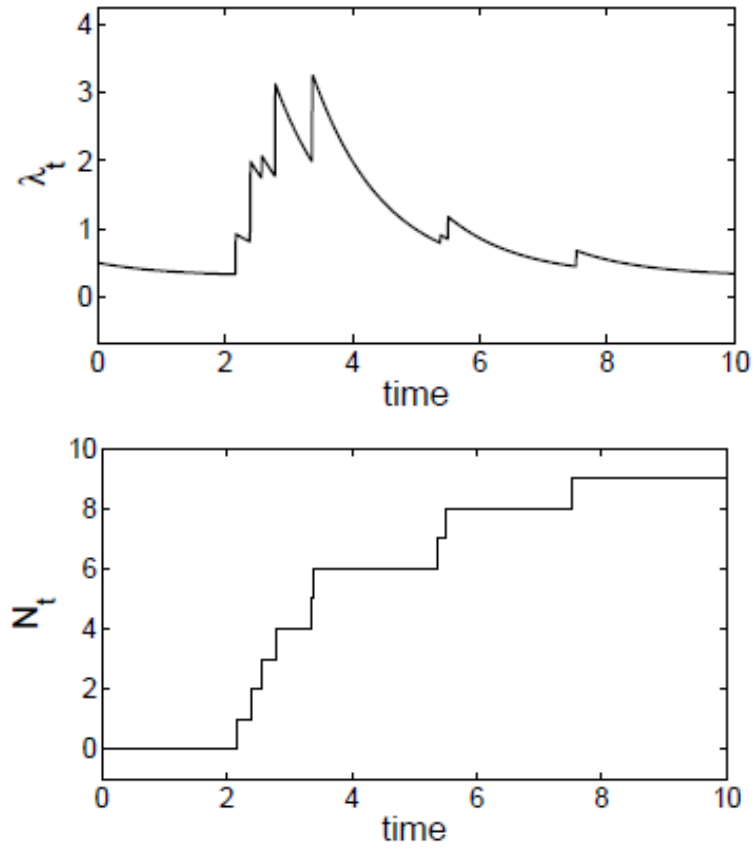
- Make the intensity λ into a deterministic function $\lambda(t)$ of time t.

  Q: What is the compensator of a PP with intensity $\lambda(t)$?
- Make the intensity an independent random process itself. This is a *Cox process*.
- Make the intensity depend on the process itself. This is a *Hawkes process* or *self-exciting process* because a jump in the process also causes the intensity to jump. In a simple version a PP N_t has intensity λ_t satisfying

$$\mathrm{d}\lambda_t = -\alpha(\lambda_t - \Lambda)\mathrm{d}t + \beta\,\mathrm{d}N_t, \qquad \alpha > 0, \quad \beta > 0.$$

  This says that a jump causes an increase in the intensity and makes another jump more likely; but the mean-reversion to the level Λ balances this.

  Q: Let $E_t^\lambda = \mathbb{E}[\lambda_t|\lambda_0]$. Find E_t^λ and find conditions on α and β under which it remains bounded as $t \to \infty$.



Top: intensity of a Hawkes Process, bottom: sample path. Note the 'clustering' of jumps (which are IID Exp(1.25)). Pictures by Xutao Kuang.

- Construct correlated Poisson processes. To construct correlated processes $N_t^{(1)}$ and $N_t^{(2)}$, take *three* independent PPs $N_t^{[i]}$, $i = 1, 2, 3$ and then set

$$N_t^{(1)} = N_t^{[1]} + \gamma N_t^{[3]},$$
$$N_t^{(2)} = N_t^{[2]} + \gamma N_t^{[3]},$$

and choose the constant γ and the intensities of $N^{[i]}$ to match the intensities of $N^{(i)}$ and the correlation coefficient (NB this is not a unique decomposition).

To match coefficients, you want $\mathbb{E}[N_t^{(i)}] = \lambda_i t$ and $\mathrm{cov}[N_t^{(1)}, N_t^{(2)}] = \rho \lambda_1 \lambda_2 t$ where ρ is the correlation coefficient. The easy way to work this out is to note

$$\mathbb{E}[N_t^{(1)} N_t^{(2)}] = \mathbb{E}[\mathbb{E}[N_t^{(1)} N_t^{(2)}] | N_t^{[3]}].$$

Levy processes

Note that the Poisson process can be defined by saying that it is the process N such that

1. $N_0 = 0$
2. $N_{t+s} - N_t$ is independent of N_t
3. $N_{t+s} - N_t$ has the Poisson distribution $\mathrm{Po}(\lambda s)$

We can ask the question - what is the class of processes that have stationary and independent increments?
This is the class of Levy processes - Brownian motion and the Poisson process are simple examples.

Markov chains

Poisson processes allow us to build continuous time Markov chains.
Discrete-time finite Markov chain: X_n has a state space labelled $1, 2, \ldots, N$. At each time step, if X_n is in state i, the transition probability for a move to state j is

$$p_{ij} = \mathbb{P}[X_{n+1} = j | X_n = i], \quad 1 \leq j \leq N.$$

Continuous time finite Markov chain: Y_t also has N states. Now define transition rates λ_{ij} for each pair of states i, j. In state i the chain waits for an exponential time $\sum_{k \neq i} \lambda_{ik}$ before moving to state j with probability

$$p_{ij} = \frac{\lambda_{ij}}{\sum_k \lambda_{ik}}.$$

The simple queue

A simple queue is a continuous time Markov chain. It can be described by X_t, the number of customers in the queue (including the one being served), at time t. For the $M/M/1$ queue we have

- Arrivals occur as a Poisson process of rate λ (M is for memoryless)
- Services occur as a Poisson process of rate μ (M)
- A single server (1)

If the arrival rate is less than the service rate, $\lambda < \mu$, the queue is stable in that the size does not grow indefinitely and it has an equilibrium distribution which is a geometric distribution with parameter $1 - \lambda/\mu$.

Queues and order books

A simple way of using queueing for order books is to consider a collection of queues with one at each tick. We can think of the system between price changes so that queue i is i ticks from the best bid or best ask.

- The arrivals to each queue are the limit orders
- The services at queues away from the best queue are the cancellations
- At the best queue the services are the cancellations and the market orders
- A price change can be thought of as occuring when the queue at either the best bid or best ask is depleted - this corresponds to the end of the busy period for that queue.

After a price change then we can reinitialize the queues and begin the process again.

Heavy traffic

The critical parameter for a simple queue is the traffic intensity $\rho = \lambda/\mu$. For stability we require $\rho < 1$.
A well established result in queueing theory is that when the queues are close to critical, that is the arrival rate λ and μ are close, in particular as $\rho \to 1$, then we can take a scaling limit.
To see how this works we take a sequence of queues X_t^n with rates given by

- $\lambda_n = \xi n$
- $\mu_n = \xi(n + c\sqrt{n})$

Now consider the rescaling of queue size by $\sqrt{n}$, in that

$$Y_t^n = X_t^n/\sqrt{n},$$

then $Y_t^n \to Y_t$ (weakly) where Y_t is a reflected Brownian motion with drift c.

Order books and heavy traffic

In order books where most of the activity takes place at the best bid and ask queues Cont and Larrard developed a heavy traffic model for the front end of the order book. The queues can be analysed analytically and by comparing with data it is shown that the arrival rates and cancellation rates are high and close.

Thus the heavy traffic limit theorem for the queueing model for an order book may be appropriate and some of the analysis of key quantities such as probability of upward or downward prices changes and time between changes can be computed using Brownian motion.

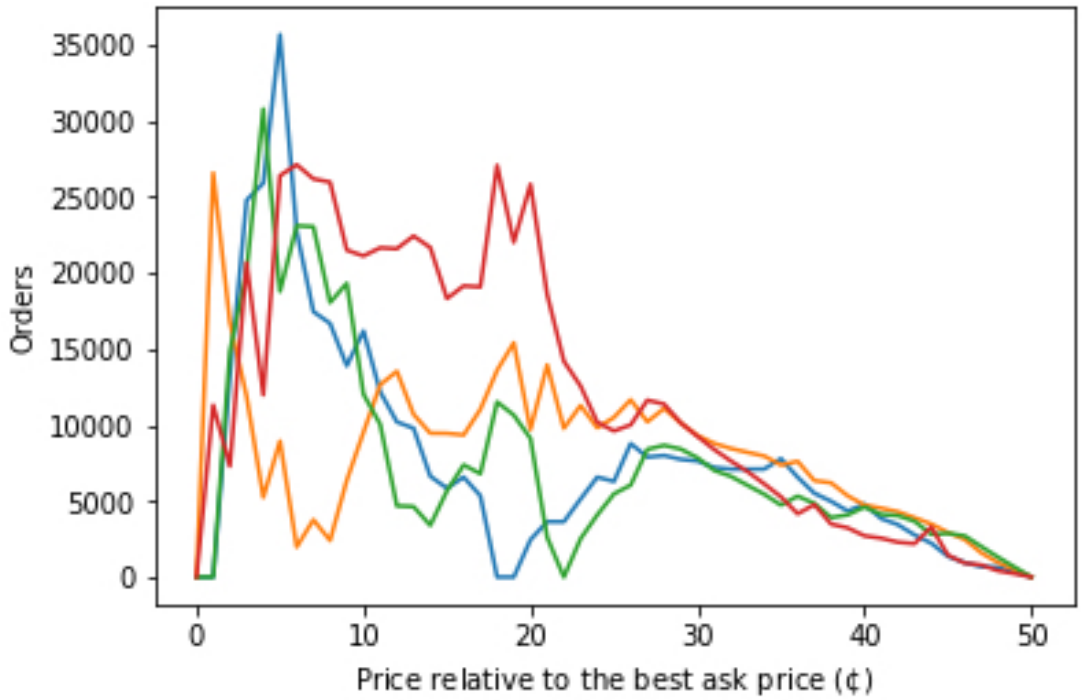
Modelling the whole book

By taking queues at each tick and analysing the arrival and cancellation rates at the ticks relative to the best price we can make a model for all the queues at once. If we assume that they are not entirely independent - people place orders in ways which tend to smooth the shape - we can build a system of SDEs for the book.

To go further take a limit as the tick size tends to 0; this gives a stochastic partial differential equation model for one side of the book! This is an equation for $u(t,x)$, the volume of orders at time t at distance x from the best price, for example

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2} + f(x) + \sigma(x)\dot{W},$$

where f represents the drift at distance x, α is a constant and $\dot{W}$ is a space-time white noise with volatiity function σ.



Some order book snapshots from an SPDE simulation.