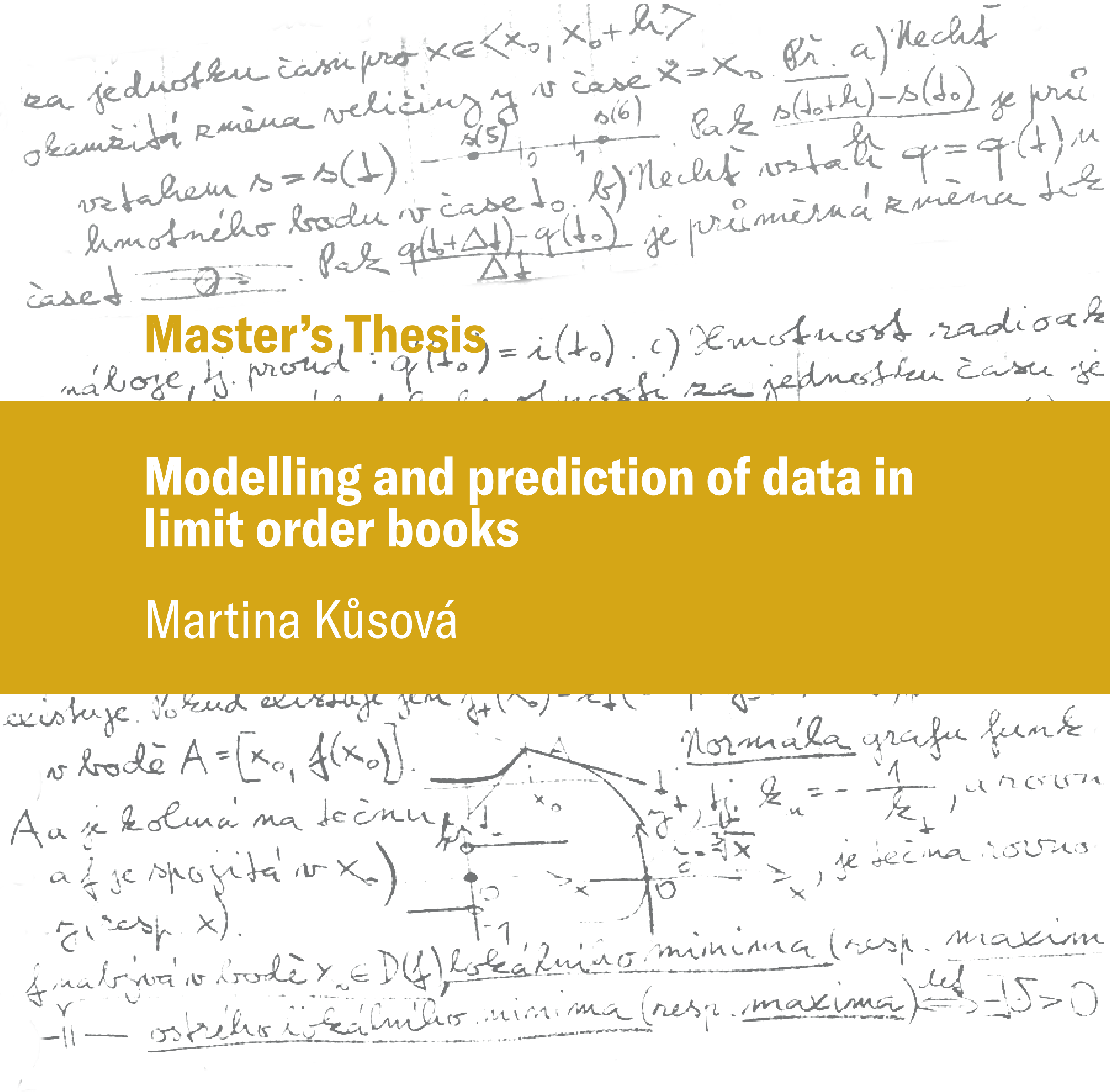
FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

DEPARTMENT OF
MATHEMATICS



Master’s Thesis

Modelling and prediction of data in limit order books

Martina Kůsová

PILSEN, CZECH REPUBLIC

2023

FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

DEPARTMENT OF
MATHEMATICS

Master's Thesis

Modelling and prediction of data in limit order books

Bc. Martina Kůsová

Thesis advisor
Ing. Jan Pospíšil, Ph.D.

PILSEN, CZECH REPUBLIC 2023

Citation in the bibliography/reference list:
KŮSOVÁ, Martina. *Modelling and prediction of data in limit order books*. Pilsen, Czech Republic, 2023. Master's Thesis. University of West Bohemia, Faculty of Applied Sciences, Department of Mathematics. Thesis advisor Ing. Jan Pospíšil, Ph.D.

ZÁPADOČESKÁ UNIVERZITA V PLZNI
Fakulta aplikovaných věd
Akademický rok: 2022/2023

ZADÁNÍ DIPLOMOVÉ PRÁCE

(projektu, uměleckého díla, uměleckého výkonu)

Jméno a příjmení:	**Bc. Martina KŮSOVÁ**
Osobní číslo:	**A21N0006P**
Studijní program:	**N0541A170005 Matematika a finanční studia**
Téma práce:	**Modelování a predikce dat v knihách limitních objednávek**
Zadávající katedra:	**Katedra matematiky**

Zásady pro vypracování

1. Zpracujte rešerši obvykle používaných modelů popisující knihy nákupních a prodejních příkazů (tzv. limit orders books).
2. Popište a ve vhodném vývojovém prostředí implementujte algoritmus kompletní rekonstrukce nákupní knihy z jednotlivých zpráv přicházejících do systému.
3. Analyzujte matematické vlastnosti vybraných modelů a postupů popisujících dynamiku objednávkové knihy.
4. Proveďte simulaci těchto modelů a kalibraci na reálná tržní data.
5. Srovnejte jednotlivé modely z pohledu úspěšnosti predikce.

Rozsah diplomové práce: **40-80 stran**
Rozsah grafických prací: **dle potřeby**
Forma zpracování diplomové práce: **tištěná**

Seznam doporučené literatury:

- Abergel, F., Anane, M., Chakraborti, A., Jedidi, A., and Toke, I. M. (2016). Limit Order Books. Physics of society: Econophysics and sociophysics. Cambridge University Press, ISBN 9781316683040, DOI: 10.1017/CBO9781316683040.
- Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. (2013). Limit order books. Quant. Finance 13(11), 1709--1742, ISSN 1469-7688, DOI: 10.1080/14697688.2013.803148.
- Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. J. Forecast. 37(8), 852-866, ISSN 1099-131X, DOI: 10.1002/for.2543.
- Rubisov, A. D. (2015). Statistical Arbitrage Using Limit Order Book Imbalance. Master's thesis, University of Toronto, pp. xi+83. URL: https://hdl.handle.net/1807/70567.

Vedoucí diplomové práce: **Ing. Jan Pospíšil, Ph.D.**
Katedra matematiky

Datum zadání diplomové práce: **3. října 2022**
Termín odevzdání diplomové práce: **22. května 2023**

Západočeská univerzita v Plzni
* Fakulta aplikovaných věd *

L.S.

Doc. Ing. Miloš Železný, Ph.D.
děkan

Doc. Ing. Marek Brandner, Ph.D.
vedoucí katedry

V Plzni dne 3. října 2022

Declaration

I hereby declare that this Master's Thesis is completely my own work and that I used only the cited sources, literature, and other resources. This thesis has not been used to obtain another or the same academic degree.

I acknowledge that my thesis is subject to the rights and obligations arising from Act No. 121/2000 Coll., the Copyright Act as amended, in particular the fact that the University of West Bohemia has the right to conclude a licence agreement for the use of this thesis as a school work pursuant to Section 60(1) of the Copyright Act.

In Pilsen, on 19 May 2023

. . . .

Martina Kůsová

Acknowledgement

I would like to thank to my advisor Ing. Jan Pospíšil, Ph.D. for his advice, suggestions, patience, and time he spent helping me with my thesis.

Abstract

Nowadays, the majority of financial markets are order-driven. An order book is a record of all buy and sell orders for a particular financial instrument. A limit order is an instruction given by a trader to buy or sell a financial instrument at a specified price or better. This thesis is focused on the modelling of data in a limit order book (LOB). We present the algorithm for LOB construction together with the basic characteristics that can be obtained from it, such as price formation and imbalance index calculation. We show that for the high-frequency LOB data, it is difficult to find the right distribution either for the inter-arrival times of incoming messages or for the order's lifetime. We jointly model the state of the LOB imbalance index and the state of the forward price movements as a two-dimensional continuous-time Markov chain. We show how state transitions, derived empirically as the statistical arbitrage opportunities matrix, lead to the possibility of an algorithmic trading strategy.

Abstrakt

V současné době je většina finančních trhů řízena objednávkami. Limitní příkaz je pokyn obchodníka k nákupu nebo prodeji finančního nástroje za danou nebo lepší cenu. Tato práce je zaměřena na modelování údajů v knize limitních objednávek (LOB). Představíme algoritmus pro konstrukci LOB spolu se základními charakteristikami, které z něj lze získat, jako je tvorba ceny a výpočet indexu nerovnováhy. Ukážeme, že pro vysokofrekvenční LOB data je obtížné nalézt správné rozdělení časů mezi příchozími objdnávkami i časů životnosti objednávek. Stav indexu nerovnováhy LOB a stav pohybu forwardových cen modelujeme společně jako dvojrozměrný Markovův řetězec se spojitým časem. Ukážeme, jak přechody stavů, odvozené empiricky jako statistická matice arbitrážních příležitostí, vedou k možné algoritmické obchodní strategii.

Keywords

Limit Order Book • price formation • imbalance index • continuous-time Markov chain

Contents

Glossary and notation 3

1 Introduction 5

2 Preliminaries 7
- 2.1 Limit Order Book . . . 7
- 2.2 Financial securities . . . 10
  - 2.2.1 Forwards and futures . . . 10
  - 2.2.2 Bonds . . . 11
- 2.3 Probability distribution . . . 12
  - 2.3.1 Exponential distribution . . . 12
  - 2.3.2 Power distribution . . . 12
- 2.4 Kolmogorov-Smirnov test . . . 13
  - 2.4.1 One-sample Kolmogorov-Smirnov test . . . 14
  - 2.4.2 Two-sample Kolmogorov-Smirnov test . . . 15

3 Methodology 17
- 3.1 The algorithm for LOB construction . . . 17
- 3.2 Continuous-time Markov chain . . . 18
  - 3.2.1 One-dimensional CTMC . . . 18
  - 3.2.2 Two-dimensional CTMC . . . 20

4 Results 23
- 4.1 Description of the data . . . 23
- 4.2 Data preparation . . . 25
- 4.3 LOB reconstruction . . . 26
  - 4.3.1 Check of the reconstruction . . . 30
- 4.4 Test of the distribution of the times . . . 31
  - 4.4.1 Exponential distribution . . . 31
  - 4.4.2 Power distribution . . . 33
- 4.5 Test of the distribution of times in smaller time intervals . . . 34

4.5.1 Exponential distribution 34
4.5.2 Power distribution 35
4.5.3 Same distribution within intervals 35
4.6 CTMC model of LOB 39

5 Conclusion **43**

A List of files **45**

Bibliography **49**

List of Figures **51**

List of Tables **53**

Glossary and notation

- LOB: limit order book
- FGBL: futures contract for a bond issued by the German federal government
- CSV: comma separated values file format
- EOBI: Enhanced Order Book Interface
- CTMC: continuous-time Markov chain
- $\mathbb{R}$: set of real numbers
- $\mathbb{N}$: set of natural numbers
- x: order
- p_x: price of order x
- w_x: size (volume) of order x
- t_x: time of order x
- $b(t)$: best buy (bid) price at time t
- $a(t)$: best sell (ask) price at time t
- $m(t)$: mid-price at time t
- $s(t)$: bid-ask spread at time t
- $I(t)$: imbalance index at time t
- $f(x)$: probability density function (pdf)
- $F(x)$: cumulative distribution function (cdf)
- $\mathbb{E}[X]$: expected value of a random variable X
- $\mathrm{Var}[X]$: variance of a random variable X

Introduction 1

This thesis is focused on the limit order book (LOB). It is a way to process incoming orders to buy or sell a certain instrument, such as a stock or a financial derivative contract. Traders can submit these orders to specify if they want to buy or sell the instrument and what their desired prices and volumes are. This system is used by more than half of financial markets, [1, p. 1]. Therefore, LOBs are studied across multiple disciplines, such as mathematics and statistics, economics and finance or computer science, [2, p. xvii].

This area of research is also affected by the growth of algorithmic trading and high frequency trading, [2, p. 4]. Several things can be looked at, such as market making strategies, optimal execution strategies or statistical arbitrage strategies, [2, p. 4]. They can be used by speculative traders who buy and sell based on their predictions, [3, p. 3].

There are usually two approaches to LOB modelling: economics-based and physics-based. The economics-based assumes perfect rationality where the traders are trying to maximize their utility, [1, p. 6]. The order flow is viewed as static, [4, p. 3]. An example of this approach is agent-based modelling, [2, p. 45]. On the other hand, the physics-based approach assumes zero intelligence. The traders are following a set of rules without any strategy. Adding, deleting and modifying the order is viewed as a stochastic process, for example, the Poisson process, [4, p. 3]. The physics-based approach includes for example Markov chain models or models where the order book is viewed as a queuing system, [2, p. 59, 77]. There are many other approaches between these two extremes with weaker assumptions on the behaviour of the traders and order flows, [1, p. 6]. It is also possible to use machine learning methods for price prediction, such as regression models or neural networks, [3, p. 3].

The evaluation of the proposed LOB models will be performed on a sample from the extensive data set coming from Eurex[1] Enhanced Order Book Interface (EOBI) provided to us for academic and research purposes by Deutsche Börse AG[2].

[1] https://www.eurex.com

[2] https://www.deutsche-boerse.com/dbg-en/

Eurex is an international electronic exchange belonging to the Deutsche Börse AG group that primarily offers trading in European based derivatives. It is the largest European futures and options market. EOBI provides complete trade data and order book information showing every visible order and quote for the most liquid Eurex futures contracts. The provided data set contains billions of order book messages per day. Raw data has several hundreds of gigabytes per day and a one-year data set is in the order of several tens to hundreds of terabytes. Roughly the same capacity is also needed to decode and process the data.

In this thesis, data for one particular product from one day is considered. In particular, we examine a data set regarding futures for a bond issued by the German government (FGBL) from December 2, 2019. There are over 700 000 data messages containing incoming orders and requests to delete or modify them. The aim of this thesis is to do a full reconstruction of the LOB using the provided data set, to model the LOB and if possible, to find a suitable trading strategy.

The structure of the thesis is as follows. In Chapter 2, we introduce LOB, futures and bonds, and some basic terms from probability and statistics, such as exponential distribution, power distribution and the Kolmogorov-Smirnov test (one-sample and two-sample). In Chapter 3, we describe the algorithm that is used to construct LOB and a Markov chain model that is used in a trading strategy. Next, in Chapter 4, we describe the data and their processing. In particular, we present the results of the LOB reconstruction, fitting the times of orders to some probability distributions and of the considered Markov chain model. We conclude in Chapter 5.

Preliminaries 2

2.1 Limit Order Book

In this section, the basic definitions related to limit order books are described. It is assumed that we have the incoming orders in the following form

$$x = (p_x, w_x, t_x), \tag{2.1}$$

where t_x is the time when it was placed, p_x is the price, and w_x is the size (volume). If it is a sell order, w_x is positive, and in case of a buy order, w_x is negative. The order is a commitment to buy (respectively sell) $|w_x|$ units of the asset for a price smaller (respectively greater) than or equal to p_x, [1, p. 2].

The smallest possible amount that can be traded is called a lot size (sometimes denoted as σ). All sizes in incoming orders must be multiples of this lot size. The smallest possible difference between prices is a tick size π. For example, if $\pi = 0.001$, it means that the maximum price smaller than 1 can be 0.999 and all prices have to be submitted with 3 decimal places. The lot size and the tick size are called LOB's resolution parameters, [1, p. 2].

The principle of the reconstruction of the LOB algorithm is that there are orders coming one by one. Each of them is either immediately matched with another previous order (a buy order is matched with a sell order and vice versa) or it becomes an active order. That means we wait for it to be matched with some other order later or for a request to delete it. It can also be modified, i.e. the trader can change the price or the volume.

The orders that are immediately matched are called *market orders* and the waiting orders are referred to as *limit orders*. The traders can choose whether they place a market or a limit order. The market order will be matched for sure. On the other hand, limit orders can be matched at better prices than market orders, but there is a risk they will not be matched at all, [1, p. 4-5].

The markets where the traders meet via LOB are called an *order-driven market*. The traders who are placing limit orders are called *liquidity providers*, whereas those who place market orders are referred to as *liquidity takers*, [2, p. 2].

LOB $\mathcal{L}(t)$ is a set of all active orders at time t. It can be divided into two parts: all active buy (bid) orders $\mathcal{B}(t)$, where $w_x < 0$, and all active sell (ask) orders $\mathcal{A}(t)$, where $w_x > 0$.

The largest buy price at time t is called a *bid price*. It can be written in the following way

$$b(t) = \max_{x \in \mathcal{B}(t)} p_x. \tag{2.2}$$

Similarly, the smallest sell price at time t is an *ask price*

$$a(t) = \min_{x \in \mathcal{A}(t)} p_x, \tag{2.3}$$

[1, p. 3]. Values $a(t)$ and $b(t)$ are also referred to as *top of book ask* and *bid* prices, respectively. The difference between them

$$s(t) = a(t) - b(t) \tag{2.4}$$

is called a *bid-ask spread*. A *mid-price* at time t is

$$m(t) = \frac{a(t) + b(t)}{2}, \tag{2.5}$$

[1, p. 3].

LOB imbalance is a ratio between the buy (bid) and the sell (ask) price. The imbalance index can be calculated as

$$I(t) = \frac{V_b(t) - V_a(t)}{V_b(t) + V_a(t)}, \tag{2.6}$$

where $V_b(t)$ is a weighted average volume at the three greatest prices in $\mathcal{B}(t)$. Exponentially decreasing weights are used, i.e.

$$V_b(t) = \mathrm{e}^{0.5 \cdot 0} \cdot w_{b_1(t)} + \mathrm{e}^{0.5 \cdot 1} \cdot w_{b_2(t)} + \mathrm{e}^{0.5 \cdot 2} \cdot w_{b_3(t)}, \tag{2.7}$$

where $w_{b_1(t)}$, $w_{b_2(t)}$ and $w_{b_3(t)}$ are the sizes at the three best buy prices. Similarly, $V_a(t)$ is a weighted average volume at the three smallest prices in $\mathcal{A}(t)$. The imbalance index can have values from -1 to 1. When there is a bigger volume of the asset on the buy side, the imbalance index is greater than zero and when it is the opposite way, the imbalance index is negative. [4, p. 6]

The LOB is usually represented with a bar chart, as it is displayed in Figure 2.1. It shows the volume at each price of all active orders on both sides at a given time

t. On the stock exchange, they can also see the internal structure of the orders. At the same price p_x there can be more different orders with different volumes w_x. For example, in Figure 2.1 we can see that at price 101, there is a volume 10. However, it can mean that there are more different orders with volumes that sum up to 10, for example, one order with volume 3, three orders with volume 2 and one order with volume 1 as displayed in Figure 2.2. These orders are sorted by an incoming time t_x.

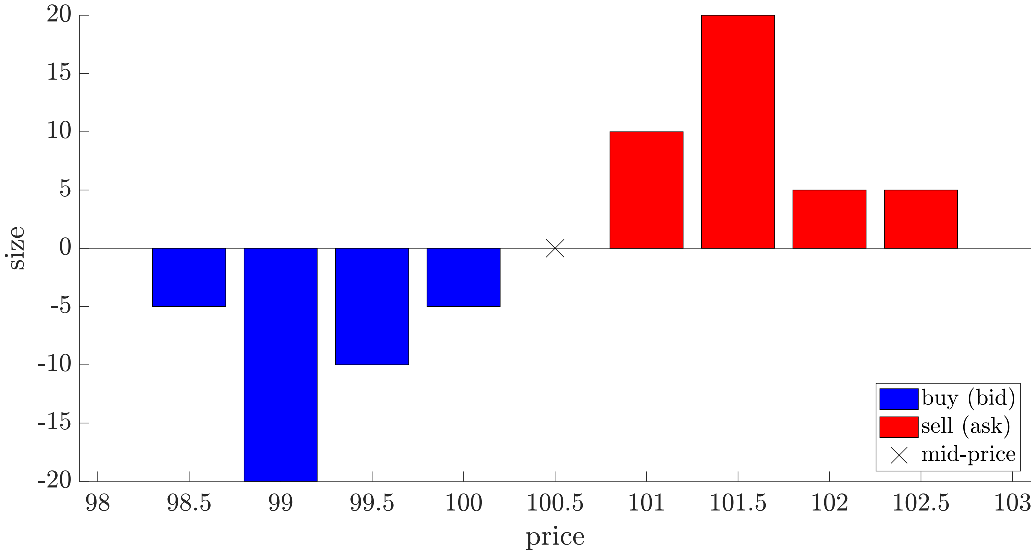


Figure 2.1: An example of LOB at time t

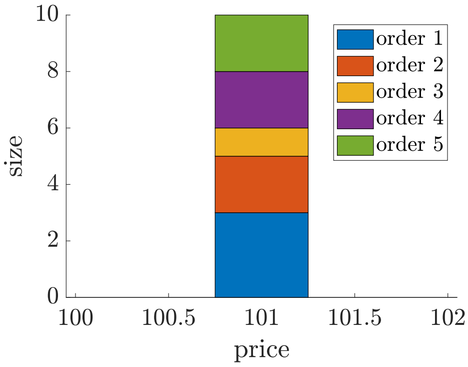


Figure 2.2: An example of orders at a certain price in LOB at time t

LOB can be also displayed in a LOBSTER format [1]. It is an $N \times 4L$ matrix, where N is the number of all orders and L is the desired number of levels. A level of LOB is the number of the best prices that will be displayed on both sides. For example, when $L = 5$, we will show the five greatest buy orders' prices and their volumes and the five smallest sell orders' prices and their sizes. This format is shown in Table

[1] https://lobsterdata.com/info/DataStructure.php

2.1. Ask Price 1 is the lowest sell price and Ask Size 1 is the size that is offered for this price. Ask Price 2 is the second lowest sell price and Ask Price 2 is the offered size for that price and so on. It is similar with Bid Price and Bid Size, where it is the highest buy price and the demanded size, [5].

Table 2.1: LOBSTER format

Ask Price 1	Ask Size 1	Bid Price 1	Bid Size 1	Ask Price 2	Ask Size 2	Bid Price 2	Bid Size 2	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

For example, a level 2 LOB for the data displayed in Figure 2.1 would look like Table 2.2.

Table 2.2: The data from Figure 2.1 displayed in LOBSTER format with level 2

Ask Price 1	Ask Size 1	Bid Price 1	Bid Size 1	Ask Price 2	Ask Size 2	Bid Price 2	Bid Size 2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
101	10	100	5	101.5	20	99.5	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.2 Financial securities

The data set in this thesis contains orders for futures of a governmental bond, so in this section, these financial securities are described.

2.2.1 Forwards and futures

The holders of forwards and futures are committed to buy or sell a certain instrument for a set price at a set time. Unlike with options, they are obligated to do so, even when it is not beneficial for them.

There are two positions: long and short. The person who must sell the underlying instrument in a given time is in a *short position*. The person who has to buy it is in a *long position*.

The main difference between forwards and futures is that futures are standardized agreements traded on organized stock exchanges, whereas forward contracts are typically privately negotiated agreements. The price of forward and future is the same if we know how the interest rate will evolve in the future. Even though we do not know this in reality, the difference between prices is not large for short-term contracts (less than nine months), [6, p. 19]. If the instrument is only weakly correlated with the interest rate, the price is similar even for long-term contracts, [6, p. 29]. Therefore, the price of forwards can also be used for futures. Another difference between them is that futures' holders (unlike forwards') in the short position pay the difference in the futures price at the end of each day when it increases or receive the amount when the price decreases. Similarly, the person in a long position obtains the difference in the futures price when it is higher and pays it when it is lower, [6, p. 26].

The underlying instrument can be for example a stock (with or without a dividend), a commodity, a currency or an interest rate. The price of a forward contract for a stock is calculated as

$$FW(t,T) = S(t)\,\mathrm{e}^{r_f(T-t)}, \tag{2.8}$$

where t is the time when the forward contract is made, T is the time when it expires, $S(t)$ is the price of the stock at time t and r_f is a continuous compound interest rate. If $FW(t,T)$ is smaller than or larger than $S(t)\,\mathrm{e}^{r_f(T-t)}$, arbitrage is possible. This price can be used for any instrument that is sold short for free, does not pay any payments during the time of the forward contract and does not have any storage or insurance expenses, [6, pp. 19–20].

In reality, the price on the market can be different since we need to include transaction expenses, short sell can be restricted or it is impossible to borrow money for a risk free interest rate, [6, p. 29].

2.2.2 Bonds

Bonds are financial securities that are connected to the debt of the issuer to the holder. When the bond is issued, the holder pays for it (the borrowed amount), then during the time before maturity the holder receives payments (coupons), and at the maturity day, the holder has a right to ask the issuer to pay the nominal value of the bond.

There are a lot of ways to divide different types of bonds, such as by time to maturity or by the issuer. The issuer can be for example the state, cities and towns, banks and companies, [7, pp. 214–216].

Every bond needs to have set a nominal value. That is the amount that will be paid to the bond holder at the time of maturity. Besides that, there is also a market value that is affected by the demand and supply at the market. The theoretical price of a bond can be calculated as a sum of the present values of the coupon payments and the present value of the nominal value of the bond. In other words, the price is

$$P = \frac{C}{1+i} + \frac{C}{(1+i)^2} + \dots + \frac{C}{(1+i)^n} + \frac{NV}{(1+i)^n}, \tag{2.9}$$

where C is a yearly coupon payment, NV is the nominal value of the bond, i is the yearly interest rate, and n is the time to maturity in years, [7, pp. 217–218].

The holder receives two different yields: the coupons and the difference between the nominal value at the time of maturity and the price at the time when he purchased the bond. Some bonds can be sold before their time to maturity. In that case, the yield is the difference between the selling price and the buying price. There are different ways to measure the yield, such as the coupon payment as a percentage of the nominal value or of the current market value, [7, p. 223].

2.3 Probability distribution

In this thesis, we attempt to fit the inter-arrival times and the times orders spent in the system to two probability distributions: exponential and power.

2.3.1 Exponential distribution

The fact that a random variable X comes from an exponential distribution with a parameter $\lambda > 0$ is written as $X \sim Exp(\lambda)$. [2] The probability density function (pdf) is

$$f(x) = \begin{cases} \lambda\, \mathrm{e}^{-\lambda x} & x \leq 0 \\ 0 & x < 0 \end{cases}. \tag{2.10}$$

The cumulative distribution function (cdf) is

$$F(x) = \begin{cases} 1 - \mathrm{e}^{-\lambda x} & x \leq 0 \\ 0 & x < 0 \end{cases}. \tag{2.11}$$

In Figure 2.3, examples of pdf and cdf with different values of λ are shown. The expected value of the exponential distribution is

$$\mathbb{E}[X] = \frac{1}{\lambda} \tag{2.12}$$

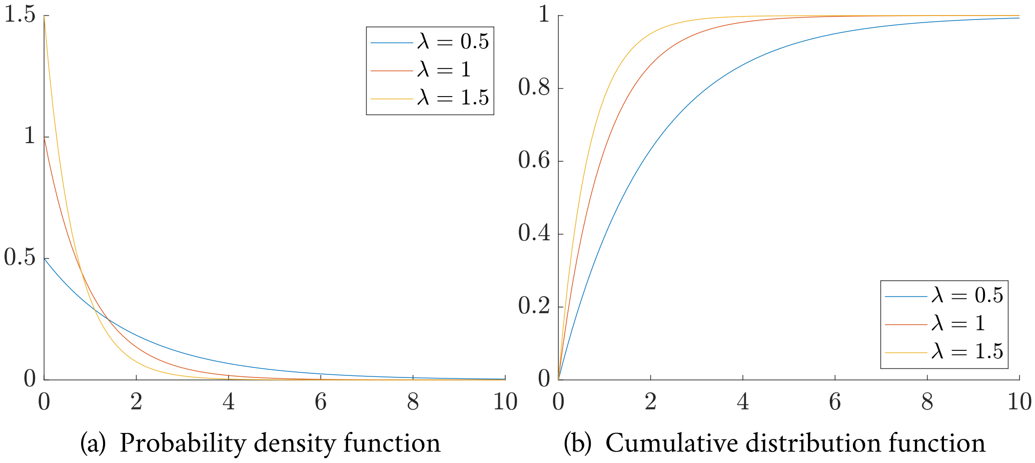
and the variance is

$$\mathrm{Var}[X] = \frac{1}{\lambda^2}. \tag{2.13}$$

2.3.2 Power distribution

Power distribution has two parameters: $a > 0$ and $k > 0$. The following definitions are from [8]. The probability density function (pdf) is

[2] Sometimes the exponential distribution is parameterized using the reciprocal $1/\lambda$.



(a) Probability density function

(b) Cumulative distribution function

Figure 2.3: Exponential distribution with different values of parameter λ

$$f(x) = \begin{cases} ak^a x^{a-1} & 0 < x \leq \frac{1}{k} \\ 0 & \text{otherwise} \end{cases}. \tag{2.14}$$

The cumulative distribution function (cdf) is

$$F(x) = \begin{cases} 0 & x \leq 0 \\ (kx)^a & 0 < x \leq \frac{1}{k} \\ 1 & x > \frac{1}{k} \end{cases}. \tag{2.15}$$

Examples of pdf and cdf with different values of parameters k and a are shown in Figures 2.4 and 2.5, respectively. The expected value is

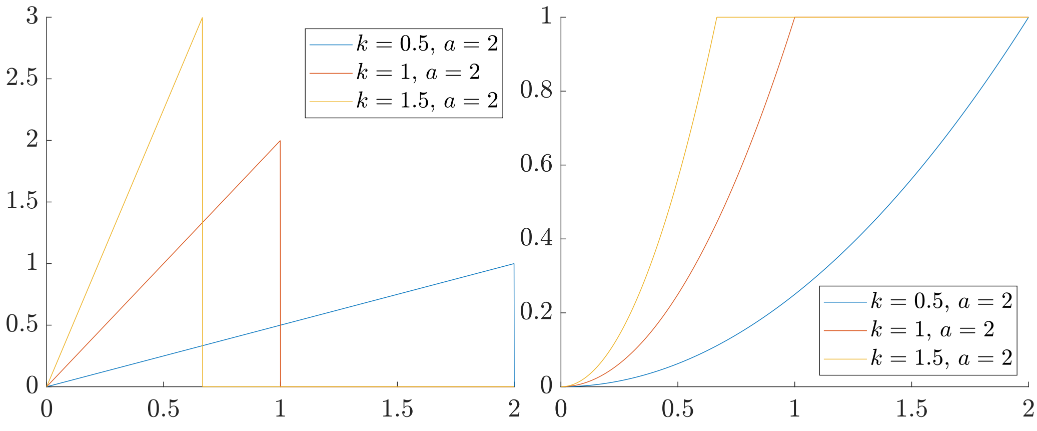
$$\mathbb{E}[X] = \frac{a}{k(a+1)} \tag{2.16}$$

and the variance is

$$\operatorname{Var}[X] = \frac{a}{(1+a)^2(2+a)k^2}. \tag{2.17}$$

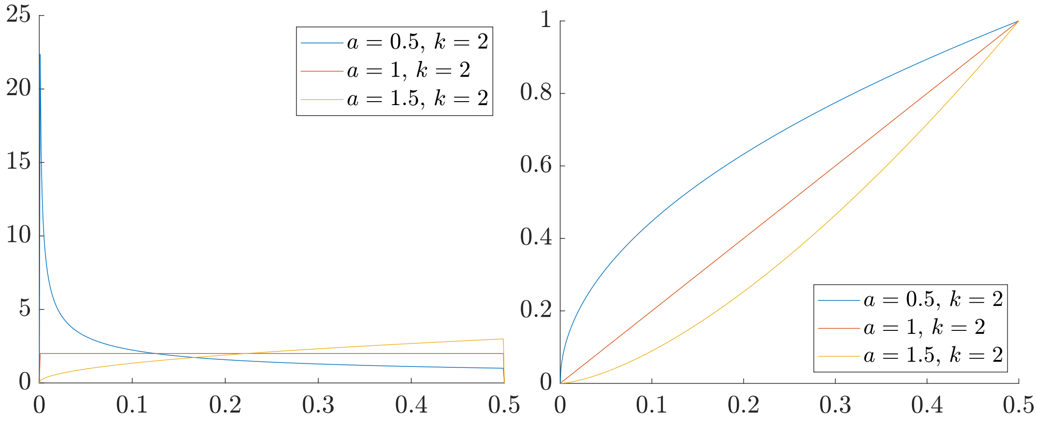
2.4 Kolmogorov-Smirnov test

Kolmogorov-Smirnov test can be either one-sample or two-sample. The one-sample test is used to test the null hypothesis that the data come from a certain distribution (for example, Gaussian or exponential), whereas the two-sample test is used when we test the null hypothesis that two data sets are from the same distribution. This section is based on [9, p. 63-65, 162].



(a) Probability density function (b) Cumulative distribution function

Figure 2.4: Power distribution with different values of parameter k



(a) Probability density function (b) Cumulative distribution function

Figure 2.5: Power distribution with different values of parameter a

2.4.1 One-sample Kolmogorov-Smirnov test

One-sample Kolmogorov-Smirnov test is a goodness of fit test. We assume we have a random sample from some unknown distribution. We also assume it is ordered, therefore, we have a sample $x_{(i)}$, where $i = 1, 2, ..., n$. This test compares the cumulative distribution function of the tested distribution with the empirical distribution function, which is defined as

$$F_n(x) = \frac{\text{the number of values in the sample that are} \leq x}{n}. \tag{2.18}$$

The test statistic is defined as

$$D = \sup\{|F_n(x) - F(x)|;\ x \in \mathbb{R}\}. \tag{2.19}$$

It is the largest distance between the values of the empirical distribution function and the cumulative distribution function of the hypothetical distribution. The test statistic can also be written as

$$D = \frac{1}{2n} + \max\left\{\left|\frac{i-0.5}{n} - F(x_{(i)})\right|;\ i = 1, 2, \ldots, n\right\}. \tag{2.20}$$

The null hypothesis that the random sample comes from a distribution with a cumulative distribution function is rejected if the D statistic is large. For a large n and a significance level α, the critical value is $\sqrt{-\frac{1}{2n}\ln(\alpha/2)}$. For $\alpha = 0.05$, the critical value is approximately $1.358/\sqrt{n}$.

If we do not know the parameters of the hypothetical distribution, we have to estimate them. The test is then called Lilliefors. The test statistic is the same but with different critical values. For an exponential distribution, $n > 30$ and $\alpha = 0.05$, the critical value is $1.06/\sqrt{n}$ (the whole table of critical values can be found in [10, p. 388]).

2.4.2 Two-sample Kolmogorov-Smirnov test

In this test, we assume that we have a random sample $X_1, X_2, ..., X_m$ from a distribution of a random variable X and a random sample $Y_1, Y_2, ..., Y_n$ from a distribution of a random variable Y. Further, we assume that these random samples are independent.

The test statistic is

$$D_{m,n} = \max_x |F_{X,m}(x) - F_{Y,n}(x)|, \tag{2.21}$$

where $F_{X,m}$ and $F_{Y,n}$ are empirical distribution functions of the two random samples. The null hypothesis that they come from the same distribution is rejected for large values of the test statistic. The table of critical values can be found in [11, p. 333-373].

Methodology 3

3.1 The algorithm for LOB construction

The algorithm for LOB construction is described in the same way as in [1, p. 5]. There are new orders coming one by one. We assume there are three types of requests (messages): requests to add a buy or sell order, requests to delete an existing order, and requests to modify an existing order.

When a new buy order $x = (p_x, w_x, t_x)$ comes at time t, there are three possibilities (the algorithm works in a similar way when a sell order comes):

1. The price of the new order is smaller than or equal to the highest buy price ($p_x \leq b(t)$). The order becomes active upon arrival. There is no change in $a(t)$ or $b(t)$.

2. If the price is greater than the highest buy price and smaller than the lowest sell price ($b(t) < p_x < a(t)$), the new order becomes active upon arrival, too. The best sell price $b(t)$ is increased to p_x.

3. The price is greater than or equal to the smallest sell price ($a(t) \leq p_x$). The new order is instantly matched to one or more waiting sell orders that do not have to be the same price as the new order. This is due to the fact that the buy order x is matched to the sell order y with the highest priority (i.e. the lowest price). If $w_x > w_y$, the remaining amount $w_x - w_y$ is matched to the sell order with the second highest priority. If there is still any amount left, it is matched to the sell order with the third priority and so on. The new lowest sell price $a(t)$ is equal to the price of the sell order with the highest priority from those that are still remaining in LOB after the matching occurred.
It can also happen that the order is executed only partially because there is not enough volume on the sell size. The remaining amount that could not be matched becomes active same as in steps 1 or 2.

In the third case, the order with the highest priority is always the one with the best price. However, when there are more orders at the same price, it must be determined what the priority of orders within that price is. The most common is price-time. That means the order with the earliest time t_x has the highest priority.

There can also be a request to delete an order because the buyer/seller does not want to wait for the order to be matched for some reason. In this case, that order is found and the size is removed from that price in LOB. It can cause a change of $a(t)$ or $b(t)$ if the deleted order had the best buy or sell price and there is no volume left at that price in LOB.

Besides that, the traders can also modify the orders. It principally works the same way as if an order with the old price and size was deleted and an order with the new price and the new size was added.

3.2 Continuous-time Markov chain

In this section, we describe definitions regarding the continuous-time Markov chain (CTMC) that is used to model the LOB imbalance index and price change later. In this thesis, we use a two-dimensional CTMC that is transformed into a one-dimensional CTMC.

3.2.1 One-dimensional CTMC

A continuous-time stochastic process $\{X(t) \mid t \geq 0\}$ [1] with a finite or countable state space K, so that $X(t) \in K$, is a *continuous-time Markov chain* if it has the *Markov property*. A Markov property means that

$$P[X(t) = j \mid X(s) = i, X(t_{n-1}) = i_{n-1}, \ldots, X(t_1) = i_1] = P[X(t) = j \mid X(s) = i], \tag{3.1}$$

where for any integer $n \geq 1$, $i_1, \ldots, i_{n-1}, i, j \in K$ are any $n + 1$ states and $0 \leq t_1 \leq \cdots \leq t_{n-1} \leq s \leq t$ is a non-decreasing sequence of $n+1$ times, [4, p. 9]. In other words, the Markov property means that the transition to the next state depends only on the current state and not on any of the previous states.

A CTMC $\{X(t) \mid t \geq 0\}$ is *time homogeneous* if for any $s \leq t$ and any states $i, j \in K$

$$P[X(t) = j \mid X(s) = i] = P[X(t - s) = j \mid X(0) = i], \tag{3.2}$$

[1] A stochastic process is a collection of random variables indexed by time. In a continuous-time stochastic process, the time indices are continuous.

[4, p. 9]. In other words, it means that the probability of transition from state i to state j is the same at any time.

The *transition rates (intensities)* $q_{i,j}$ specify the rate (intensity) at which X jumps from state i to j, , [4, p. 9]. Furthermore, we can define the *conditional transition probabilities* $p_{i,j}$, i.e. the probability that the state jumps to state j conditional on currently being in state i, [4, p. 9]. The amount of time that X spends in state i between entering and jumping to another state is exponentially distributed with the rate v_i. It is called a *holding time*, [4, p. 9]. They have the following relations

$$v_i = \sum_{j \in K, j \neq i} q_{i,j}, \tag{3.3}$$

$$q_{i,j} = v_i \cdot p_{i,j} \tag{3.4}$$

and

$$p_{i,j} = \frac{q_{i,j}}{v_i}, \tag{3.5}$$

[4, p. 9].

A *(transition) intensity matrix* (or *infinitesimal generator matrix*) G of a CTMC $\{X(t) \mid t \geq 0\}$ is defined as

$$g_{i,j} = \begin{cases} q_{i,j} & i \neq j \\ -v_i & i = j \end{cases}, \tag{3.6}$$

[4, p. 9].

If we denote the transition probabilities of a CTMC

$$P_{i,j}(t) = P[X(t) = j \mid X(0) = i] \tag{3.7}$$

and matrix $P(t) = P_{i,j}(t)$, then

$$P(t) = GP(t) \tag{3.8}$$

and

$$P(t) = \mathrm{e}^{Gt}, \tag{3.9}$$

[4, p. 9], where

$$\mathrm{e}^{Gt} = \sum_{j=0}^{+\infty} \frac{G^j t^j}{j!} \tag{3.10}$$

is a matrix exponential.

3.2.2 Two-dimensional CTMC

The two-dimensional CTMC is used to model the imbalance index and the price change. The imbalance index I is very erratic, so it is smoothed first using a mean of the values in a certain time interval in the past $[\Delta t_I, t]$, [4, p. 12]. Another possibility is to do a moving average, i.e. the mean of a certain number of past values and the current value, [12]. Since the imbalance index can have real values from an interval $[-1, 1]$, the smoothed imbalance index is discretized into subintervals (bins), [4, p. 8]. The bin at time t is denoted as $\rho(t)$. It can have values $1, 2, \ldots, B$, where B is the number of bins.

The price change we look at is the mid-price change $\Delta m(t)$. It is calculated as a signum function of the change of the price in a future time interval $[t, \Delta t_m]$

$$\Delta m(t) = \mathrm{sgn}(m(t + \Delta t_m) - m(t)), \tag{3.11}$$

[4, p. 12]. Again, if we want to get a price change at the current time, it is also possible to select some number, for example 20, and then calculate the difference between the price twenty time steps from the current time and the price in the current time and get the signum of this value, [12]. The price change can have values -1, 0, or 1. It is equal to -1 when the price decreases, 0 when it stays the same, and 1 when it increases.

The two-dimensional CTMC consists of $\rho(t)$ and $\Delta m(t)$. We can encode it to a one-dimensional CTMC using an encoding function $\varphi(\rho(t), \Delta m(t))$. An example of such encoding, where we assume 3 bins for the imbalance index, is shown in Table 3.1.

Table 3.1: An example of encoding with 3 bins for imbalance index

$\varphi(\rho(t), \Delta m(t))$	$\rho(t)$	$\Delta m(t)$
1	1	-1
2	2	-1
3	3	-1
4	1	0
5	2	0
6	3	0
7	1	1
8	2	1
9	3	1

We cannot know the state of the CTMC in real time, because $\Delta m(t)$ contains the future price change. It is possible to estimate the intensity matrix G by the maximum likelihood method. The estimate of the transition intensities (non-diagonal elements of matrix G) is

$$\hat{q}_{i,j} = \frac{N_{i,j}(T)}{H_i(T)}, \tag{3.12}$$

where $N_{i,j}(T)$ is the number of transitions from state i to j in time interval $[0, T]$ and $H_i(T)$ is the holding time in state i in time interval $[0, T]$, [4, p. 11]. In other words, this means that the estimate of transition intensities from state i to j is the number of transitions from state i to j divided by the total time spent in state i, [4, p. 11]. We can calculate the diagonal elements of matrix G using Equation (3.3)

$$\hat{v}_i = \sum_{j \in K, j \neq i} \hat{q}_{i,j}, \tag{3.13}$$

[12].

The estimated matrix G can be transformed into a one step transition probability matrix

$$P = \mathrm{e}^{G \Delta t_I}, \tag{3.14}$$

where the elements of this matrix are

$$P_{i,j} = P[\varphi(\rho_{[t-\Delta t_I, t]}, \Delta m_{[t, t+\Delta t_m]}) = j \mid \varphi(\rho_{[t-2\Delta t_I, t-\Delta t_I]}, \Delta m_{[t-\Delta t_I, t]}) = i], \tag{3.15}$$

[4, p. 16]. In other words, it means

$$P_{i,j} = P[\varphi(\rho_{\text{curr}}, \Delta m_{\text{future}}) = j \mid \varphi(\rho_{\text{prev}}, \Delta m_{\text{curr}}) = i] \tag{3.16}$$

or after rewriting it to two dimensions

$$P_{i,j} = P[\rho_{\text{curr}} = i, \Delta m_{\text{future}} = j \mid \rho_{\text{prev}} = k, \Delta m_{\text{curr}} = m], \tag{3.17}$$

[4, p. 16]. We can use this formula in an equation

$$\begin{gathered} P[\Delta m_{\text{future}} = j \mid \rho_{\text{prev}} = k, \Delta m_{\text{curr}} = m, \rho_{\text{curr}} = i] = \\ = \frac{P[\rho_{\text{curr}} = i, \Delta m_{\text{future}} = j \mid \rho_{\text{prev}} = k, \Delta m_{\text{curr}} = m]}{P[\rho_{\text{curr}} = i \mid \rho_{\text{prev}} = k, \Delta m_{\text{curr}} = m]} \end{gathered} \tag{3.18}$$

where the nominator is the same as Equation (3.17) and the denominator can be calculated as

$$\begin{gathered} P[\rho_{\text{curr}} = i \mid \rho_{\text{prev}} = k, \Delta m_{\text{curr}} = m] = \\ = \sum_{b} P[\rho_{\text{curr}} = i, \Delta m_{\text{future}} = j \mid \rho_{\text{prev}} = k, \Delta m_{\text{curr}} = m], \end{gathered} \tag{3.19}$$

[4, p. 16]. This means we can calculate the probability that the price will transition in the future to state j knowing the current price change, the current bin of the imbalance index and the previous bin of the imbalance index. All these probabilities can be stored in a matrix denoted by Q. The size of this matrix is $3B \times 3B$.

Several trading strategies can be done based on matrix Q, for example, a Naive Trading Strategy. In this strategy, a buy (respectively sell) order is executed if the probability of an upward (respectively downward) price change is greater than 0.5, [4, p. 17].

Results 4

4.1 Description of the data

The data set in this thesis was provided for academic and research purposes by Deutsche Börse AG. We had data with a futures contract for a bond issued by the German federal government (FGBL) from December 2, 2019. These contracts are traded on Eurex Exchange. [1] Eurex Enhanced Order Book Interface (EOBI) gives complete information about trades and LOB, that show all visible orders and prices of the most liquid futures contracts on the Eurex exchange.

A complete data description is available in the EOBI manual [13]. The prices are in EUR. All these data sets include information (EOBI messages) available in comma separated value (CSV) file format. The following files were available:

- `OrderAdd`: incoming orders,
- `OrderDelete`: requests to delete orders,
- `OrderModify`: requests to change orders,
- `OrderModifySamePrio`: requests to change orders but keeping the same time priority timestamp,
- `OrderMassDelete`: requests to empty the whole LOB and
- other files that are not used in this thesis.

`OrderAdd` looks like the data in Table 4.1. It has the following columns:

- `PARENT_ID`,
- `ID`,

[1] Eurex is an international exchange that offers mainly European derivatives. It is the largest European market with options and futures contracts.

- TrdRegTSTimeIn: time of the order,
- SecurityID: unique identificator of the instrument,
- TrdRegTSTimePriority: priority timestamp,
- DisplayQty: size,
- Side: 1 meaning buy and 2 meaning sell and
- Price: offered/demanded price,

[13, p. 51].

Table 4.1: A sample of order add data

PARENT_ID	ID	TrdRegTSTimeIn	SecurityID	TrdRegTSTimePriority	DisplayQty	Side	Price
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1452493	5311	1575245704137166516	4128839	1575245454393025894	0.0004	2	172.99
1452493	5312	1575245704137166516	4128839	1575245542983775817	0.0004	1	172.93
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The same columns are in OrderDelete with one additional column:

- TransactTime,

[13, p. 58].

In OrderModify there is PARENT_ID, ID, TrdRegTSTimeIn, SecurityID and, Side which are the same as in OrderAdd and OrderDelete and these columns:

- TrdRegTSPrevTimePriority: previous priority timestamp,
- PrevPrice: previous price,
- PrevDisplayQty: previous size,
- TrdRegTSTimePriority: new priority timestamp,
- DisplayQty: new size,
- Price: new price,

[13, p. 54].

OrderModifySamePrio contains columns PARENT_ID, ID, TrdRegTSTimeIn, TransactTime, SecurityID, Side which are the same as in OrderAdd and OrderDelete and these columns:

- PrevDisplayQty: previous size,
- TrdRegTSTimePriority: priority timestamp
- DisplayQty: new size,
- Price: price of the order,

[13, p. 56].

To do the whole reconstruction of the LOB, OrderMassDelete should be included, too. It has columns: PARENT_ID, Security_ID and TransactTime which are the same as in the other files, [13, p. 60]. In the given data set, there is only one row that corresponds to deleting the LOB at the end of the day because the TransactTime is December 2, 2019, 21:00:00, [14]. Therefore, this file was not used in the reconstruction of LOB.

All the timestamps are in the so-called epoch time, in particular, it is in nanoseconds from 00:00:00 UTC on January 1, 1970, [13, p. 24]. Each order has a unique key consisting of the instrument identifier, the priority timestamp and the side, i.e. the columns SecurityID, TrdRegTSTimePriority and Side, [13, p. 51].

4.2 Data preparation

Firstly, the data were prepared in order to do the reconstruction of LOB later. In Table 4.2 we can see how many rows are in each of the order files. In total, there are 735 795 of them. This data should contain only records with one SecurityID (an identification of the instrument), so the rows where this number was incorrect were deleted. This inconsistency was caused by a temporarily incorrect parsing of provided large EOBI data sets (several hundreds of gigabytes per day). The process of parsing EOBI data was beyond the scope of the thesis and we worked with already parsed data provided in CSV format. The numbers of rows that were deleted from each file are also shown in Table 4.2 as well as the number of rows that were left after deleting. In total, 2 929 rows were deleted and 732 866 were left.

Table 4.2: Number of messages in the FGBL 2. 12. 2019 order files

File	Number of messages		
	before deleting	deleted	after deleting
OrderAdd	316 295	540	315 755
OrderDelete	285 016	379	284 637
OrderModify	119 626	1 797	117 829
OrderModifySamePrio	14 858	213	14 645
total	735 795	2 929	732 866

All columns containing the size (DisplayQty and PrevDisplayQty) were multiplied by 1000 in all files in order to have the sizes as integers.

The prices and sizes in time were plotted. The TrdRegTSTimeIn column was used as a time. The plots are displayed in Figures 4.1, 4.2, 4.3, and 4.4. Even though we can see there are some extreme values in prices and sizes that could be incorrect, no more rows were deleted, because the complete data set is needed to do the reconstruction of the LOB later. The lower or higher prices at the beginning of the day in OrderAdd could be caused by the initialization of the LOB. The LOB is usually set up to some initial state by adding some buy and sell requests at the beginning of the day before the traders start placing orders. Another reason for extreme values can be speculation. The traders can place buy orders at a lower price because they hope the prices will go down and they will make a profit, or they want to sell it at a higher price because they hope the price will go up.

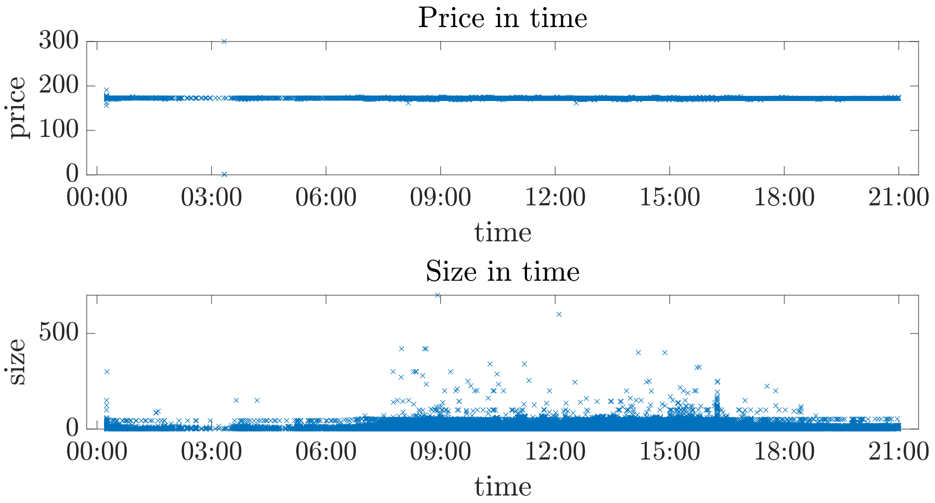


Figure 4.1: The price and size of FGBL in 2. 12. 2019 in time in OrderAdd file

A new column was added to the data. In OrderAdd it was filled with ones, in OrderDelete with twos and in both OrderModify and OrderModifySamePrio with threes, so we can recognize what type of request is in the row later. All these data were put into one table that was sorted by time.

4.3 LOB reconstruction

The prepared data were used to reconstruct LOB. The prices in the algorithm were multiplied by 100 for convenience to have them stored as integers, i.e. the same way as size, time, ID etc. (in the following figures and tables, they are divided by 100 back).

In every step, the best sell price $a(t)$ and the best buy price $b(t)$ are calculated (if there are any). The mid-price $m(t)$ and spread $s(t)$ are also computed. If there are at

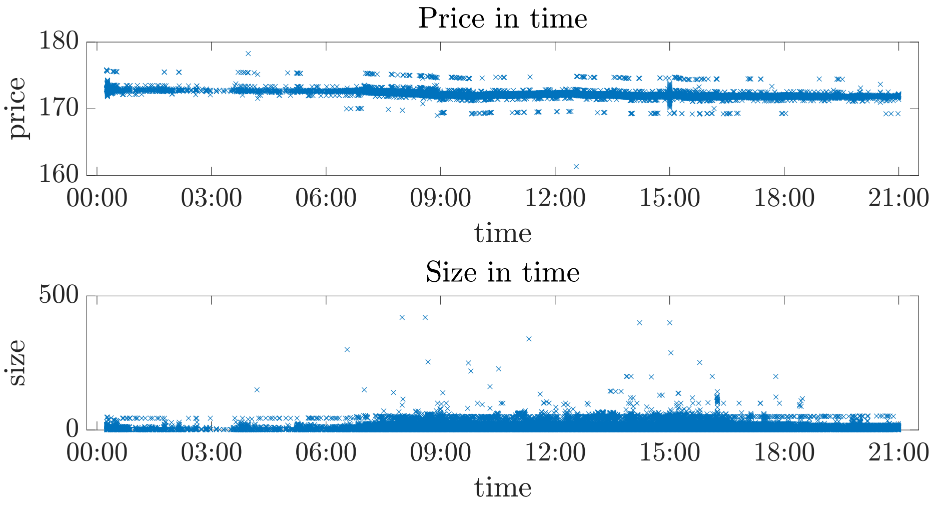


Figure 4.2: The price and size of FGBL in 2. 12. 2019 in time in `OrderDelete` file

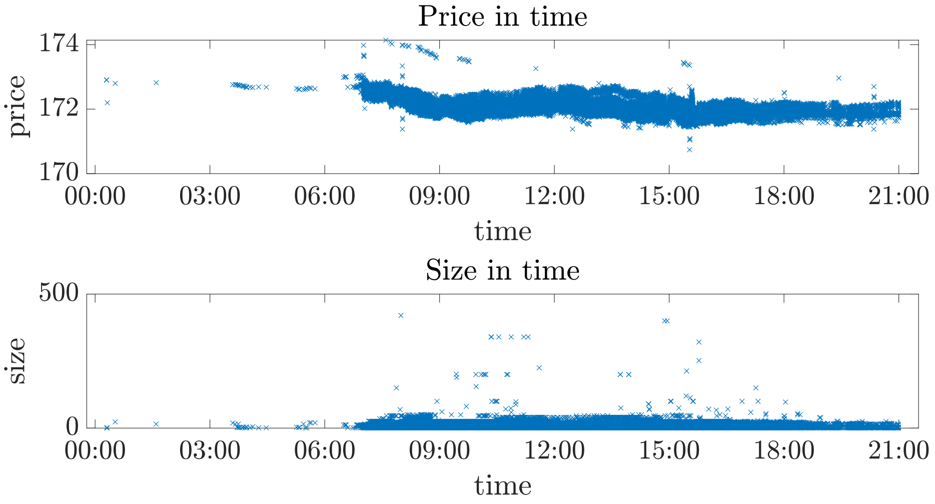


Figure 4.3: The price and size of FGBL in 2. 12. 2019 in time in `OrderModify` file

least three different prices on each side, the imbalance index $I(t)$ is calculated using (2.6).

Every time a match of orders occurs, the time when the waiting order was matched is saved with its ID, so we can test the distribution of the times that the orders spent in the system later.

After all steps, the code saves the LOB in a LOBSTER format in a CSV file. It can be chosen what level of LOB we want to save and display.

In Figure 4.5 there is a LOB at 10:00. A level 5 LOB at that time is shown in Table 4.3. The highest buy price was 171.97 with size 123 and the lowest sell price was 171.98 with size 21. Therefore, the mid-price was 171.975 and the spread was 0.01. The imbalance index was 0.3151. It is positive, which means there is a bigger volume on the buy side.

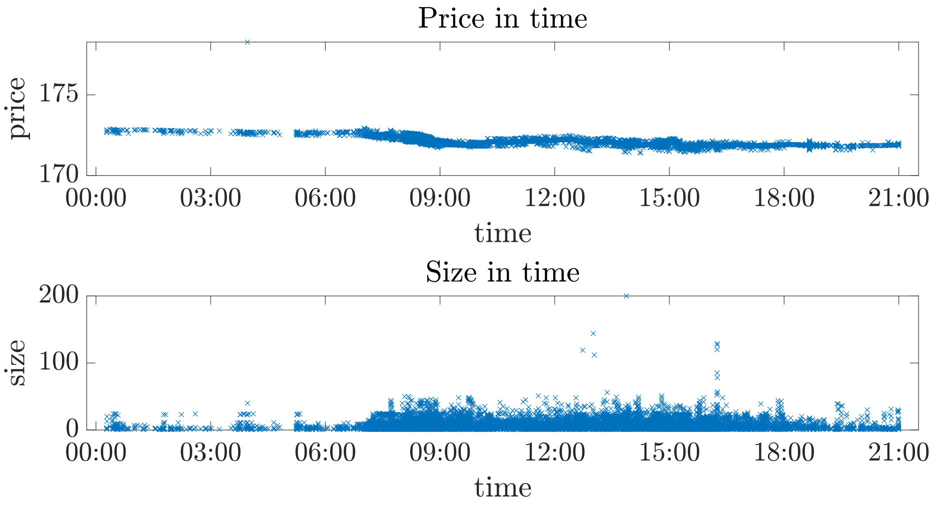


Figure 4.4: The price and size of FGBL in 2. 12. 2019 in time in `OrderModifySamePrio` file

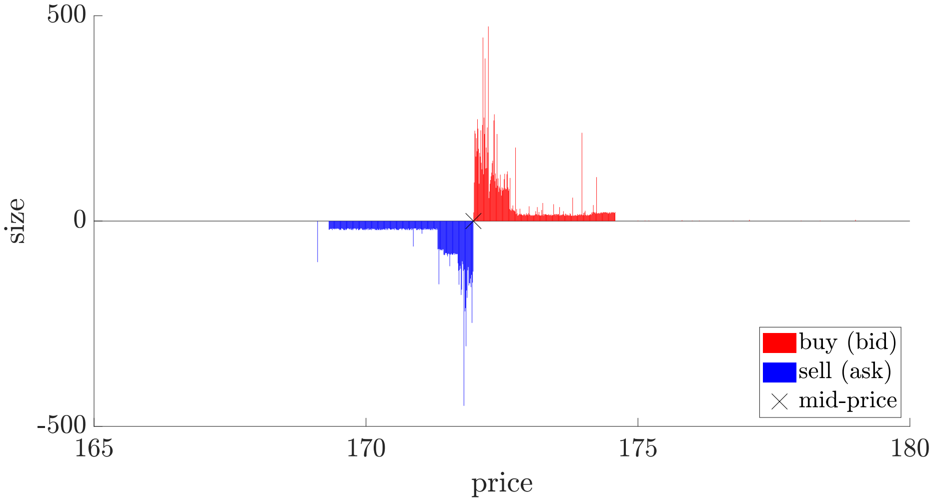


Figure 4.5: The LOB at 10:00

Table 4.3: Level 5 LOB at 10:00

	Buy		Sell	
	Price	Size	Price	Size
1	171.97	123	171.98	21
2	171.96	150	171.99	94
3	171.95	248	172.00	220
4	171.94	131	172.01	213
5	171.93	161	172.02	157

The LOB at the end of the day is displayed in Figure 4.6 and the level 5 LOB is shown in Table 4.4. The best buy price was 171.86 and the best sell price was

171.89, so they both decreased from the morning, as we can also see in Figure 4.7, where it is shown how both best prices evolved in time. The mid-price was 171.875, the spread was 0.03 and the imbalance index was -0.1026. It is negative, therefore, there is a bigger size on the sell size. In Figure 4.8, it is shown how the mid-price, the spread and the imbalance changed during the day. In the attached file `FGBL_2019_lev5_orderbook.csv` in the `data-out` folder, we can find the complete level 5 LOB in LOBSTER format and in the file `FGBL_2019_lev30_orderbook.csv`, there is a level 30 LOB in LOBSTER format.

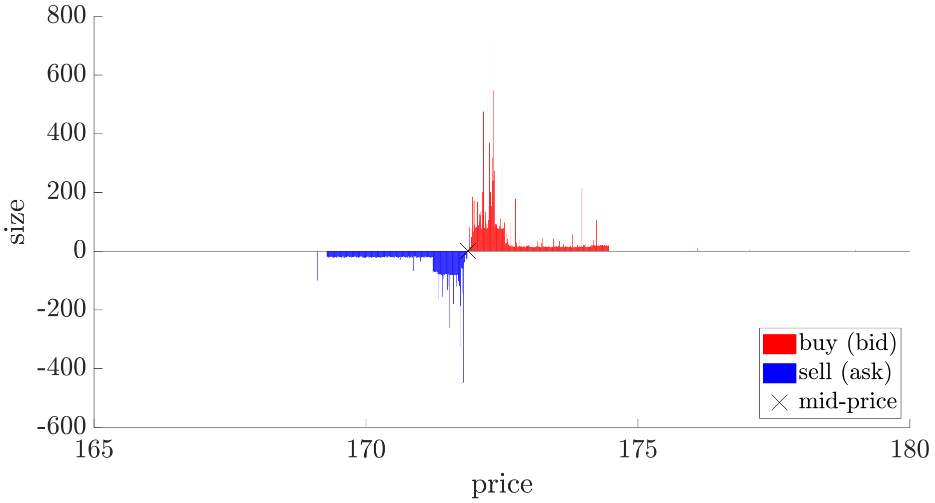


Figure 4.6: The LOB at the end of the day

Table 4.4: Level 5 LOB at the end of the day

	Buy		Sell	
	Price	Size	Price	Size
1	171.86	23	171.89	8
2	171.85	27	171.90	81
3	171.84	34	171.91	18
4	171.83	13	171.92	15
5	171.82	31	171.93	26

In Figures 4.9 and 4.10, there is a zoom of the figures to the time interval 10:00-10:05 to see the changes in the prices, spread and imbalance index more precisely. It is a time when the traders have already placed some orders after initialization. In Figure 4.11, we can see a comparison of the best buy price, the best sell price and the mid-price in this time interval.

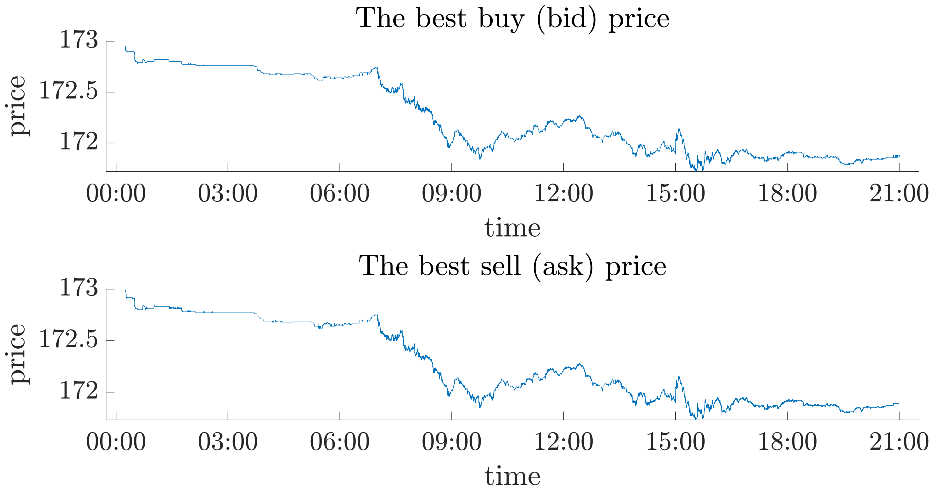


Figure 4.7: The best buy and the best sell price in time

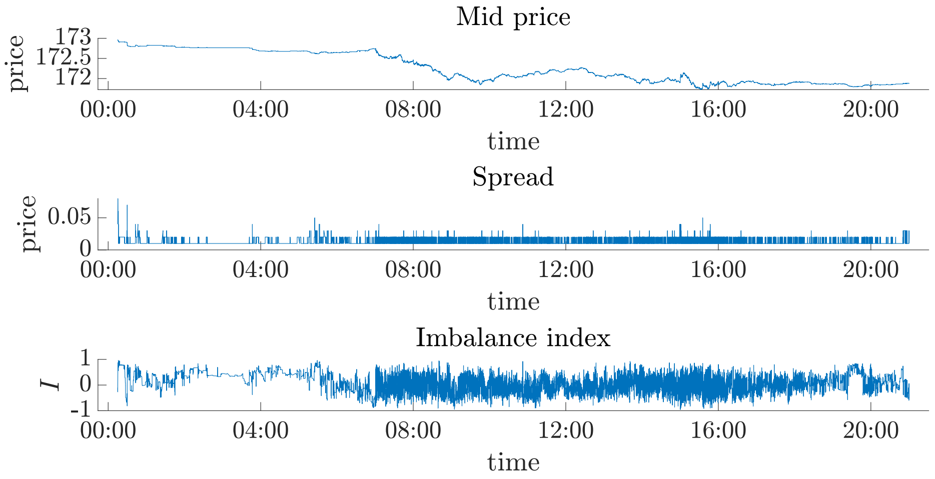


Figure 4.8: The mid-price, the spread and the imbalance index in time

4.3.1 Check of the reconstruction

There are several ways how to check if the reconstruction of the LOB was done correctly. The best way would be to check it with LOB from Deutsche Börse AG, but it was not available at the time of writing the thesis directly.

It is also possible to check the prices from LOB reconstruction with prices that can be found on the Eurex website, [15]. We can find there the open and closed price for the day and the daily lowest and highest price. However, we do not know what price it is in the LOB (if it is the mid-price or something else) or if it was not adjusted somehow. Therefore, it is not the best way to check the reconstruction.

It is also possible to do the check with `FullOrderExecution` and `PartialOrderExecution` files that were included in the data set that was available but it was beyond the scope of this thesis.

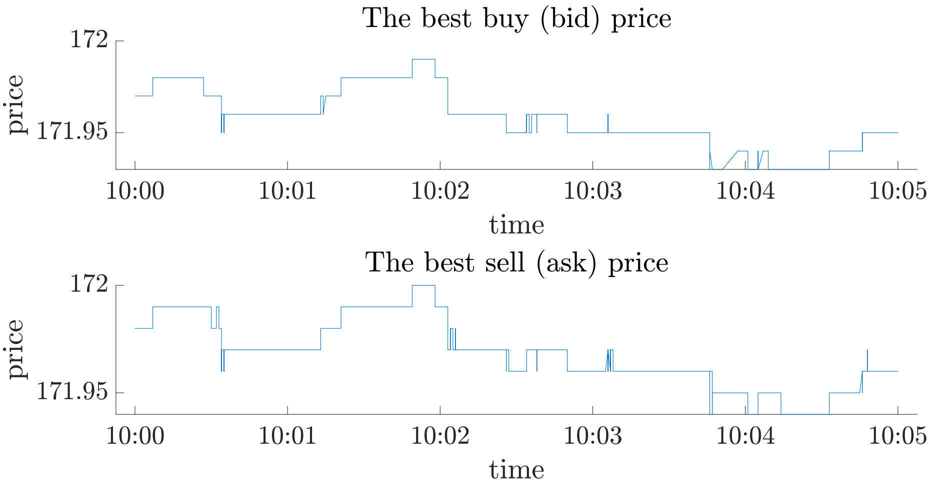


Figure 4.9: The best buy and the best sell price in time between 10:00 and 10:05

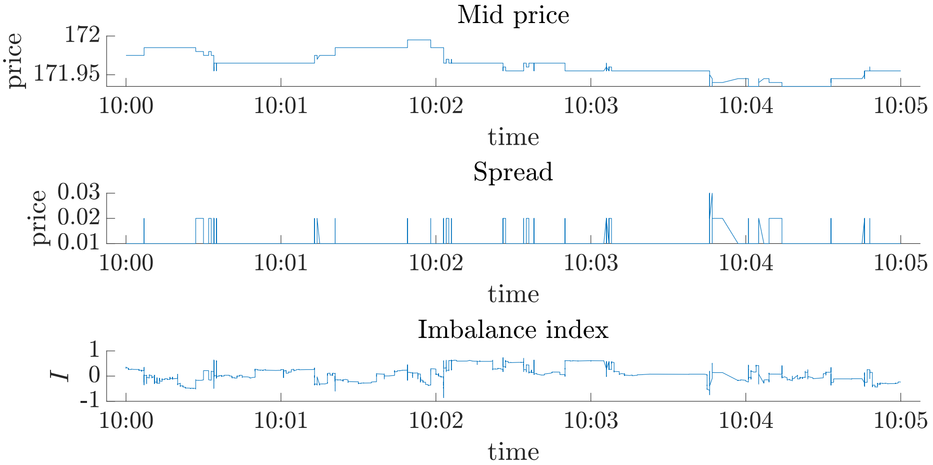


Figure 4.10: The mid-price, the spread and the imbalance index in time between 10:00 and 10:05

4.4 Test of the distribution of the times

In this section, we show what is the distribution of the times in LOB. There were two sets of times tested: the times between two consecutive incoming add orders (inter-arrival times) and the times that the orders spent in the system (the time between adding the order and matching it to another one or deleting it).

4.4.1 Exponential distribution

Firstly, it was tested whether both of these times come from an exponential distribution, see section 2.3.1. If they did, we could model the LOB dynamics as a Poisson process.

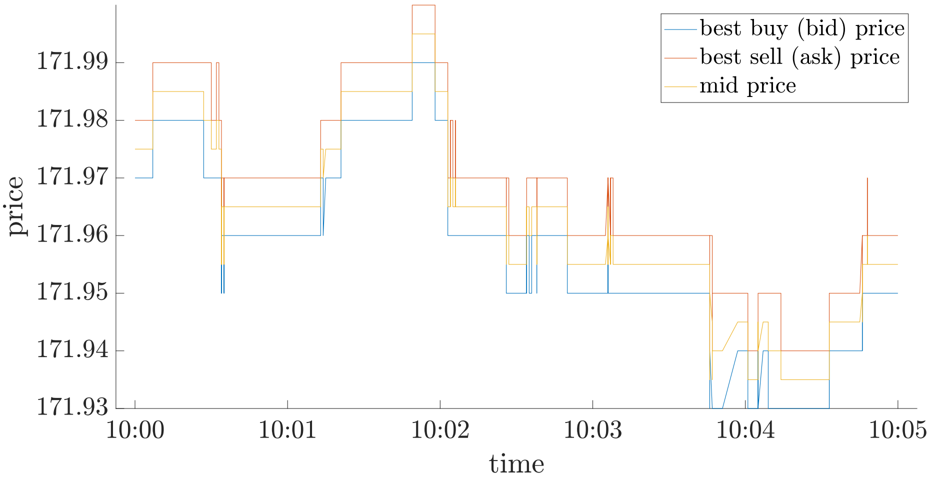


Figure 4.11: The best buy price, the best sell price and the mid-price in time between 10:00 and 10:05

4.4.1.1 The times of adding orders

The times in the `OrderAdd` file are in the so-called epoch time format, in particular, in nanoseconds from January 1, 1970, 00:00:00 UTC, so they were first adjusted to be in nanoseconds since the beginning of the day on December 2, 2019, 00:00:00 UTC. The value 1575244800×10^9 (midnight of December 2, 2019, in nanoseconds [14]) was subtracted from each time in the `TrdRegTSTimeIn` column in `OrderAdd`.

An exponential distribution was fitted to these times. The estimated parameter is $\lambda = 4.2273 \times 10^{-9}$. In Figure 4.12, there is also a QQ plot that shows the quantiles of the fitted exponential distribution vs. the quantiles of the sample. If the distributions were the same, the points would be on a diagonal line.

Even though it does not look like the times are from the exponential distribution in the graph, it was tested using a one-sample Kolmogorov-Smirnov test. The null hypothesis is that the times come from the fitted exponential distribution with $\lambda = 4.2273 \times 10^{-9}$. The alternative hypothesis is that the times do not come from the fitted exponential distribution. It was tested at a 5% significance level[2] and the result of the test was that the null hypothesis was rejected. The p-value of the test was 0.

4.4.1.2 The times between adding and matching/deleting orders

The times when the orders were added and deleted are in the incoming data sets. An intersection of the orders' IDs (`TrdRegTSTimePriority` column) was found to find orders that were added and later deleted. Then a difference between the time of deleting and the time of adding was calculated for each of these orders. Midnight did

[2] i.e. $\alpha = 0.05$, the corresponding confidence level is $1 - \alpha = 0.95$, i.e. 95%

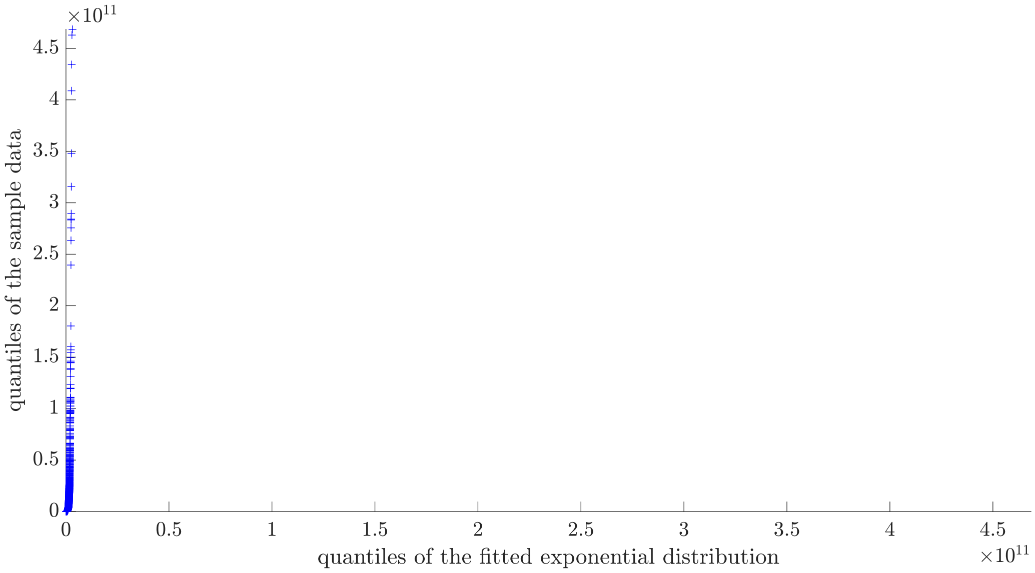


Figure 4.12: QQ plot of the quantiles of the fitted exponential distribution vs. quantiles of the times of adding orders

not have to be subtracted this time because we are only interested in the difference between the times. Similarly, the times between adding and matching the orders were calculated. The IDs of the orders that were matched and the times when it happened were saved during the LOB reconstruction. The data set that was tested in this section contains both of these times.

Same as before, first the exponential distribution was fitted to the data. The estimated parameter is $\lambda = 5.5924 \times 10^{-12}$. The null hypothesis that the times between adding and matching/deleting come from the fitted exponential distribution with the estimated parameter λ was tested with one-sample Kolmogorov-Smirnov test, too. The alternative hypothesis is that the times do not come from the fitted exponential distribution. The null hypothesis was rejected at a 5% significance level with a p-value 0.

4.4.2 Power distribution

Since it was rejected that both sets of times come from an exponential distribution, it was tested whether the data could come from a power distribution, see section 2.3.2.

4.4.2.1 The times of adding orders

Firstly, the power distribution was fitted to the times. The estimated parameters are $a = 1.5541 \times 10^7$ and $k = 3.8908 \times 10^8$. One-sample Kolmogorov-Smirnov test was used to test the null hypothesis that the data come from a fitted power distribution with the estimated parameters a and k. The alternative hypothesis was that they do not come from this fitted distribution. At a 5% level of significance, the null hypothesis was rejected. The p-value of the test was 0.

4.4.2.2 The times between adding and matching/deleting orders

The power distribution was fitted to the times that the orders are in the system, too. The estimated parameters are $a = 1.6275 \times 10^9$ and $k = 1.9207 \times 10^{11}$. Once again, Kolmogorov-Smirnov test was used to test the null hypothesis that the times between adding and matching or deleting the orders come from the fitted power distribution with the estimated parameters a and k. The null hypothesis was rejected at a 5% significance level. The p-value of the test was 0.

4.5 Test of the distribution of times in smaller time intervals

Since the exponential and power distribution was not fitted to the data, the inter-arrival times (times of adding orders) were divided into smaller time intervals (1 hour, 15 minutes, 10 minutes) and it was tried to fit the exponential or power distribution to the data in these intervals. The aim was to find some time in a day when they have a certain distribution. It was not done with the times between adding and matching/deleting the orders because it would be difficult to find a way to divide them into intervals (based on the time of adding the order, deleting/matching the order, etc.).

4.5.1 Exponential distribution

Firstly, the inter-arrival times were divided into 15 minutes intervals: 0:00-0:15, 0:15-0:30, etc. In each of these intervals, the exponential distribution was fitted and then Kolmogorov-Smirnov test was used to test the null hypothesis that the times of adding orders in a given interval come from the fitted exponential distribution with the estimated parameter λ. The alternative hypothesis was that they do not come from the fitted distribution. There was only one interval out of 96, where the null hypothesis was not rejected at a 5% significance level. It was between 3:15 and 3:30. The estimated parameter in this interval is $\lambda = 0$. In the remaining 95, intervals the null hypothesis was rejected at a 5% significance level.

The same was also tried with 10 minutes and 1 hour time intervals. We can see a comparison of all three ways in Table 4.5. In the 10 minutes intervals, only in 2 out of 144 intervals, the null hypothesis was not rejected at a 5% significance level. The intervals were 2:50-3:00 and 3:20-3:30. In case of 1 hour intervals in all of the 24 intervals, the null hypothesis was rejected at a 5% significance level.

Table 4.5: The proportion of time intervals where the null hypothesis that the times of add orders come from a fitted exponential distribution was not rejected

interval	proportion of not rejected	percentage of not rejected
10 minutes	2/144	1.39%
15 minutes	1/96	1.04%
1 hour	0/24	0.00%

4.5.2 Power distribution

Because the exponential distribution could not be fitted in most of the time intervals, it was tried to fit a power distribution. It was done with 1 hour, 15 minutes and 10 minutes time intervals as well, however, none of these intervals were fitted. The null hypothesis that the times in the given time interval come from the fitted power distribution was rejected at a 5% significance level in all of them.

4.5.3 Same distribution within intervals

Since it was not possible to fit an exponential or power distribution to most of the intervals, it was tested whether the times of add orders divided into smaller time intervals come from the same distribution.

Two-sample Kolmogorov-Smirnov test was used to do that. The null hypothesis was that the two sets of times come from the same distribution and the alternative hypothesis was that they do not come from the same distribution. All pairs of 15 minutes intervals were tested. In total, there are 4560 pairs of different time intervals. This was calculated as $\frac{96 \cdot 96 - 96}{2}$ (96 is the number of 15 minutes intervals in one day). In Figure 4.13, we can see the results of each of the tests. In 327 out of the 4560 tests, the null hypothesis was not rejected at a 5% level of significance. In the others, it was either rejected or there were not enough data in one of the intervals to do the test. This happened at the end of the day when there were fewer incoming orders. The p-values of all tests are shown in Figure 4.14. We can see that the pairs of the same interval were not rejected with a p-value 1, because they were the same data sets. Most of the other not rejected null hypotheses were in the time intervals after midnight. This could be caused by the initialization of the LOB.

Two intervals where the null hypothesis was not rejected and where both of them contained at least 100 data were chosen to plot. In Figure 4.15, we can see histograms of times of add orders in these two intervals and in Figure 4.16, a QQ plot of their quantiles. Even though the null hypothesis was not rejected, the distribution does not look very similar in the QQ plot. We can see an example of two time intervals where the null hypothesis was rejected in Figures 4.17 and 4.18.

10 minutes and 1 hour intervals were also tried. In Table 4.6 we can see the

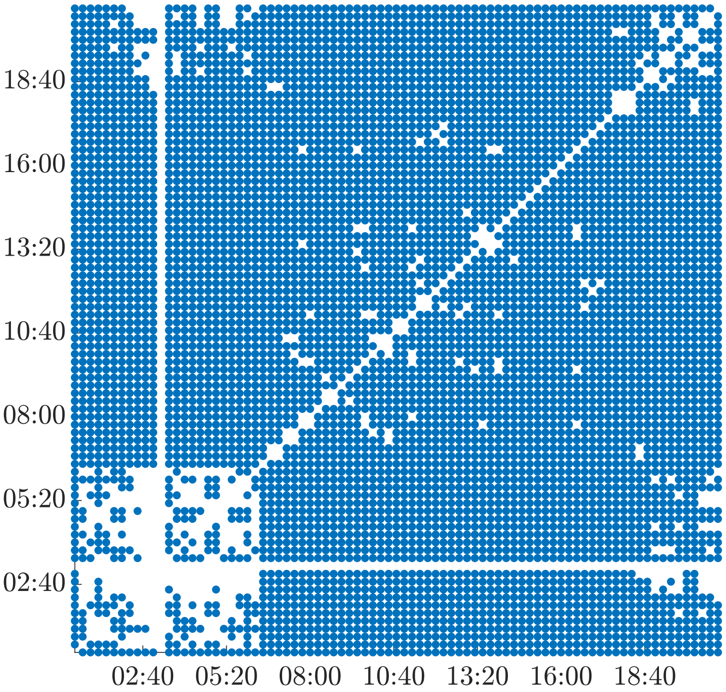


Figure 4.13: The result of a test that times of add orders divided into 15 minutes intervals come from the same distribution (blue dots mean rejecting the null hypothesis and blank spaces mean rejecting)

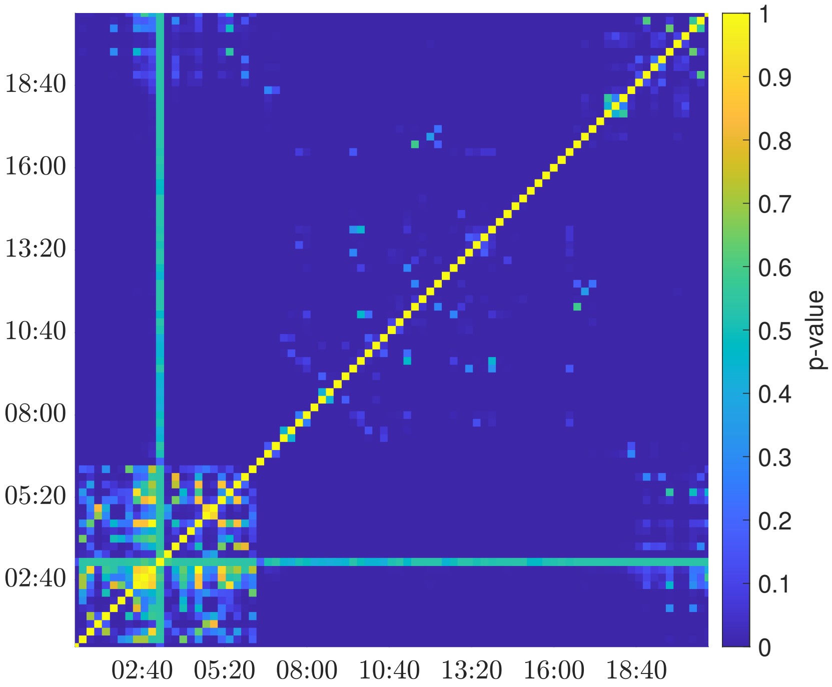
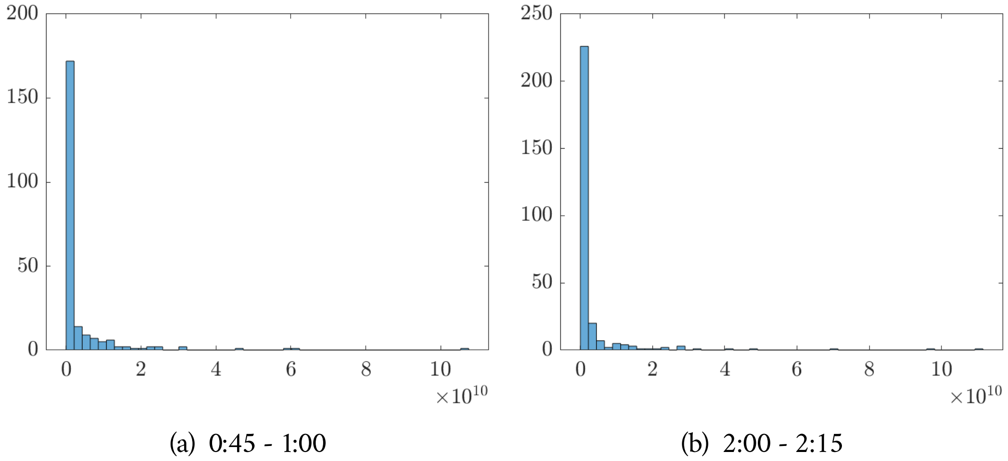


Figure 4.14: P-values of a test that times of add orders divided into 15 minutes interval come from the same distribution



(a) 0:45 - 1:00

(b) 2:00 - 2:15

Figure 4.15: Histograms of two chosen intervals where the null hypothesis was not rejected

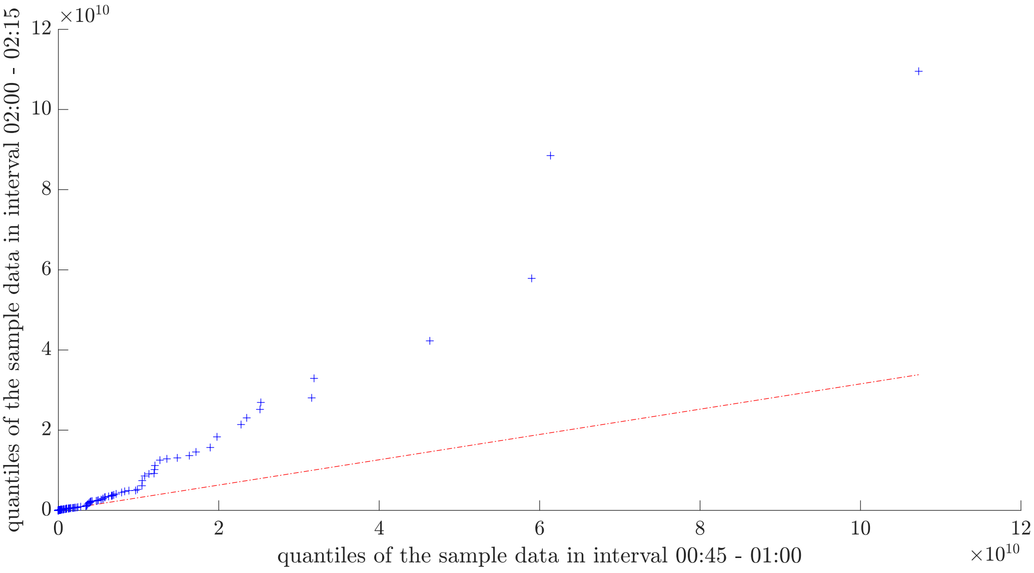
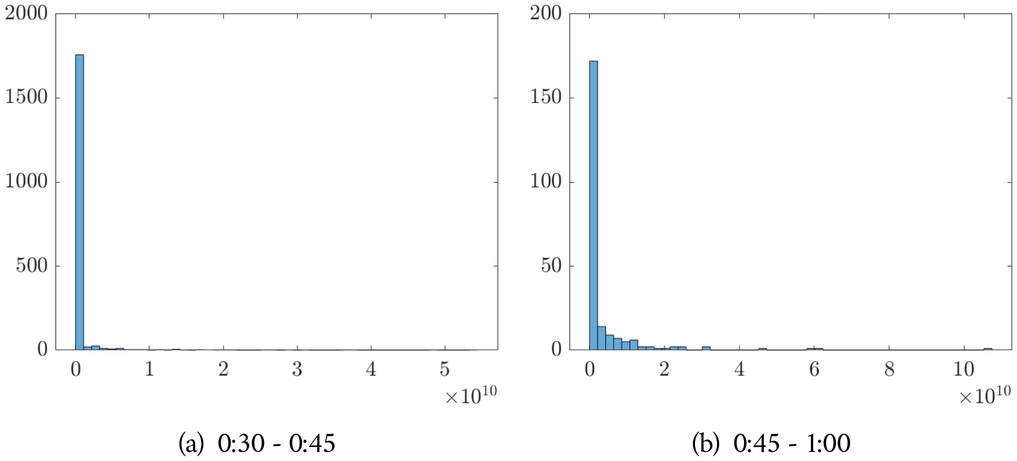


Figure 4.16: QQ plot of quantiles of data in intervals 0:45 - 1:00 and 2:00 - 2:15 where the null hypothesis was not rejected

results. The null hypothesis about the same distribution was not rejected for 742 out of 10 296 pairs of time intervals in case of 10 minutes intervals. When it comes to 1 hour, it was 8 out of 276 pairs.

Since we did not manage to fit any distribution to any of the two considered time sets, in the rest of the thesis we will focus on the imbalance index dynamics.



(a) 0:30 - 0:45

(b) 0:45 - 1:00

Figure 4.17: Histograms of two chosen intervals where the null hypothesis was rejected

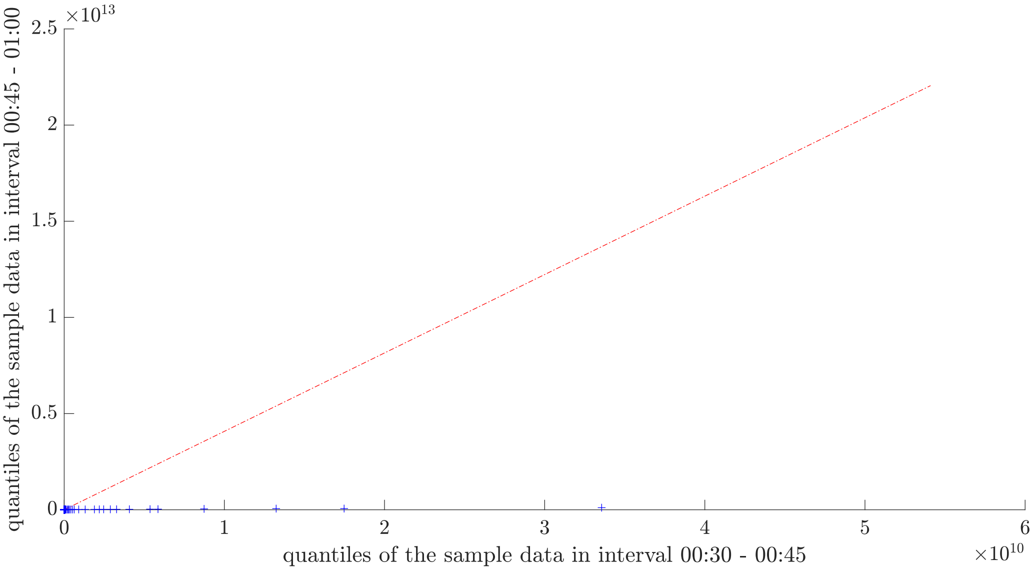


Figure 4.18: QQ plot of quantiles of data in intervals 0:30 - 0:45 and 0:45 - 1:00 where the null hypothesis was rejected

Table 4.6: The proportion of the pairs of time intervals where the null hypothesis that the times of add orders come from the same distribution was not rejected

interval	proportion of not rejected	percentage of not rejected
10 minutes	742/10296	7.21%
15 minutes	327/4560	7.17%
1 hour	8/276	2.90%

4.6 CTMC model of LOB

In this section, a two-dimensional CTMC is used to model the price change and imbalance index (see section 3.2) in the reconstructed LOB. In Figure 4.19 on the left, we can see how the mid-price and imbalance index from LOB evolved during the day (between 10:00 and 16:00 when there were more incoming orders). It was zoomed to a time interval of 10:00-10:05 to see it in detail, see Figure 4.19 on the right.

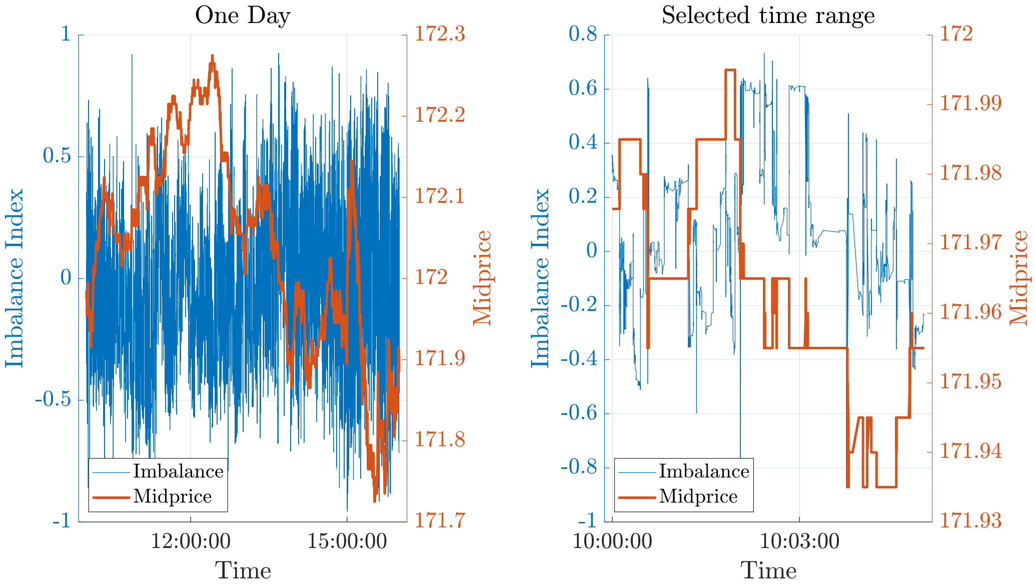


Figure 4.19: Mid-price and imbalance index in time intervals 10:00-16:00 and 10:00-10:05

Firstly, the imbalance index was smoothed. It was done using a moving average over the last 10 values. The smoothed imbalance index in both time intervals is displayed in Figure 4.20. Next, the smoothed imbalance index change was discretized into 3 bins to get the values $\rho(t)$.

We calculated the price change with a window of 20 values. For example, the first price was subtracted from the twenty-first price, the second price from the twenty-second and so on. From this value, we calculated signum to get the price change $\Delta m(t)$. In Figure 4.21, we can see what the discretized imbalance index and mid-price change look like in the time interval 10:00-16:00 and in more detail in the time interval 10:00-10:05.

Next, the values of the encoding function $\varphi(\rho(t), \Delta m(t))$ were calculated similarly as it was shown in Table 3.1 in Section 3.2.2. For the price change equal to -1 and bins of the imbalance index 1, 2, 3, the values of the encoding function were 1, 2, 3, for the price change 0 they were 4, 5, 6 and for the price change 1 they were 7, 8, 9.

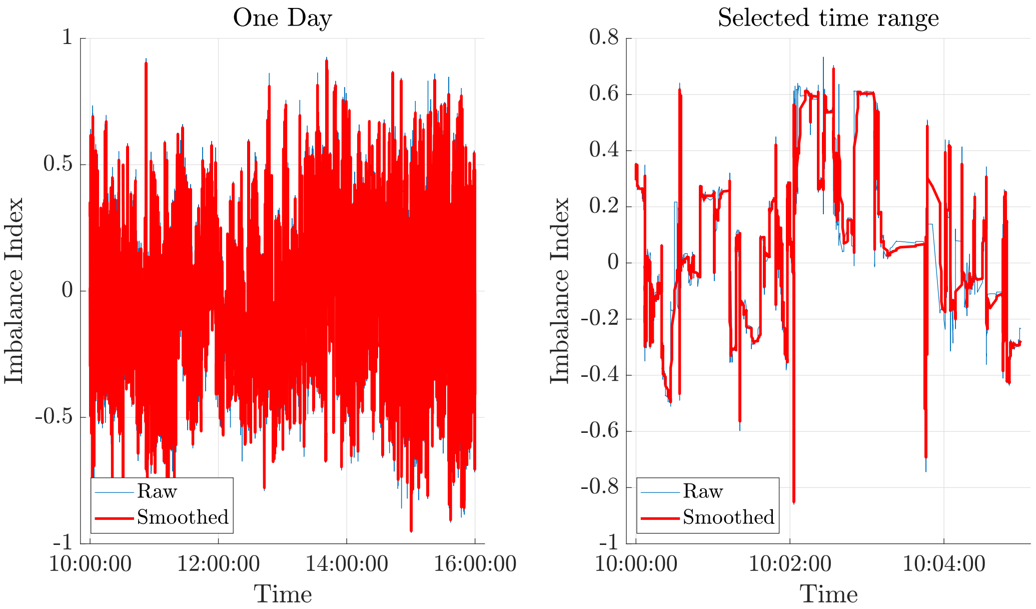


Figure 4.20: Smoothed imbalance index in time intervals 10:00-16:00 and 10:00-10:05

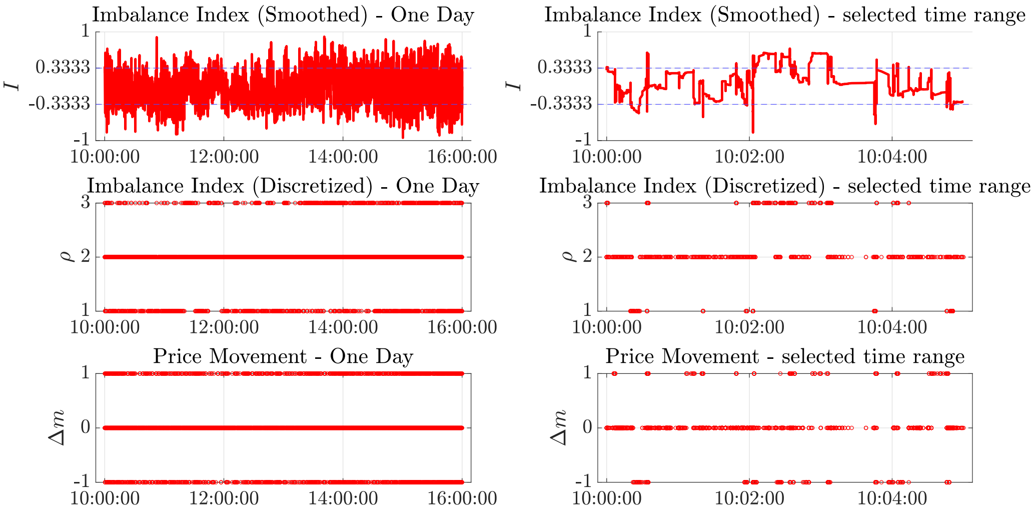


Figure 4.21: Smoothed imbalance index, discretized imbalance index and mid-price change in time intervals 10:00-16:00 and 10:00-10:05

The numbers of transitions between each pair of states were calculated. The holding times were estimated as the number of times when there was a transition from the current state to the same one, [12]. The intensity matrix G was estimated as it was described in section 3.2.2 and the one step transition probability matrix P was calculated using equation (3.14).

The probability of price changes in the future conditional on the current price change, the current bin of imbalance index and the previous bin of imbalance index were calculated using (3.18). They are saved in a matrix that is shown in Table 4.7 and they are plotted in Figure 4.22. In the middle yellow part, we can see that the situation with the highest probability is that the price will not change in the future when it did not change now. For example, when $\rho_{prev} = 1$, $\Delta m_{curr} = 0$ and $\rho_{curr} = 1$, there is a 0.9073 probability that $\Delta m_{future} = 0$. In other words, when we know that the bin of imbalance index was previously 1, now it is also 1 and the price did not change now, there is a 90.73% chance that the price will stay the same in the future. The blue parts show the least probable situations. Those are the transitions from current price change 0 and 1 to future price change -1 (the left lower part of the figure) and the transitions from current price change -1 and 0 to future price change 1 (the right upper part of the figure). For example, when $\rho_{prev} = 3$, $\Delta m_{curr} = 1$ and $\rho_{curr} = 1$, there is a 0.17 probability that $\Delta m_{future} = -1$. In other words, when we know that the bin of imbalance index was previously 3, now it is 1 and the price increased now, there is a 17% chance that the price will decrease in the future.

Table 4.7: The probabilities of the future price change conditional on the current price change, the current imbalance index bin and the previous imbalance index bin

	(1,-1)	(2,-1)	(3,-1)	(1,0)	(2,0)	(3,0)	(1,1)	(2,1)	(3,1)
(1,-1)	0.6215	0.3408	0.1129	0.3703	0.6252	0.7538	0.0082	0.0340	0.1333
(2,-1)	0.6655	0.3694	0.1280	0.3281	0.6060	0.7743	0.0064	0.0245	0.0977
(3,-1)	0.7083	0.4294	0.1470	0.2862	0.5436	0.7938	0.0055	0.0270	0.0591
(1,0)	0.0741	0.0538	0.0203	0.9073	0.8712	0.7515	0.0186	0.0750	0.2283
(2,0)	0.1117	0.0332	0.0186	0.8732	0.9336	0.8695	0.0151	0.0332	0.1120
(3,0)	0.2660	0.0867	0.0213	0.7207	0.8631	0.9046	0.0133	0.0502	0.0741
(1,1)	0.0566	0.0239	0.0088	0.7366	0.5339	0.3083	0.2068	0.4422	0.6829
(2,1)	0.1057	0.0245	0.0099	0.8074	0.6062	0.3544	0.0869	0.3693	0.6358
(3,1)	0.1700	0.0416	0.0110	0.7657	0.6319	0.3998	0.0643	0.3265	0.5892

The trading strategy is to find if there is a probability greater than 0.5 of a price increase (respectively decrease) and then we can buy (respectively sell) the asset. In Figure 4.23, it is displayed which elements of the matrix are greater than 0.5. For example, when $\rho_{prev} = 1$, $\Delta m_{curr} = -1$ and $\rho_{curr} = 1$, the probability that $\Delta m_{future} = -1$ is greater than 0.5. Therefore, we can sell the asset, when the previous imbalance index bin was 1, the current imbalance index bin is 1 and the price change currently decreased, because we expect that the price will decrease in the future, too. On the other hand, one of the situations when we would buy the asset because we expect the price to increase (with a probability higher than 50%), is when $\rho_{prev} = 3$, $\Delta m_{curr} = 1$ and $\rho_{curr} = 3$.

This trading strategy depends on the choice of parameters (the number of values in the moving average calculation, the number of bins for the imbalance index and the length of the window for price change). Therefore, the results may be different when trying different parameters.

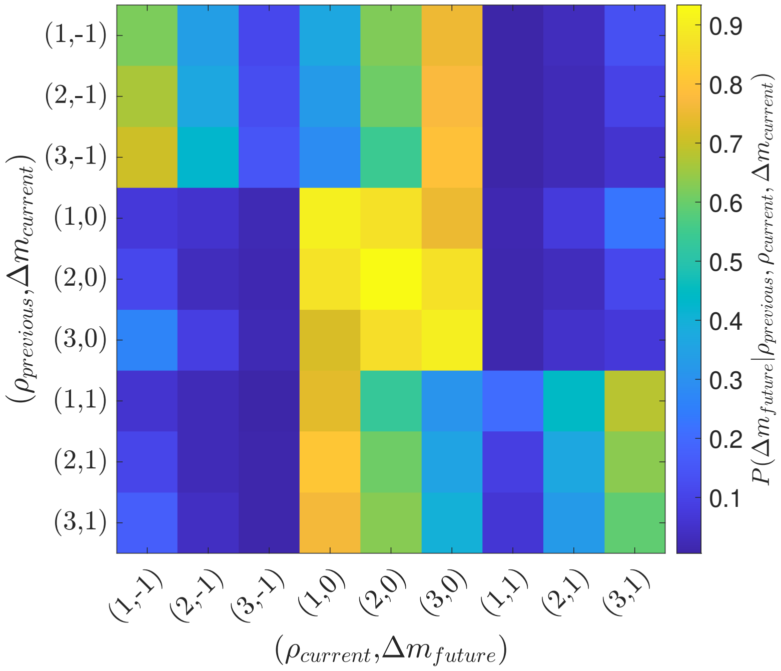


Figure 4.22: The probabilities of the future price change conditional on the current price change, the current imbalance index bin and the previous imbalance index bin

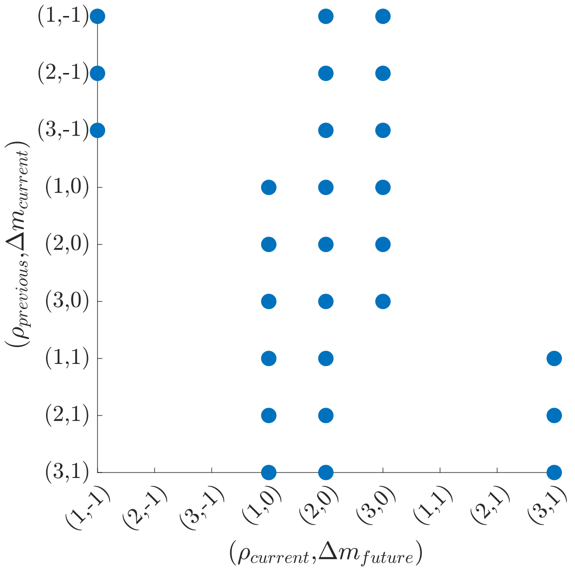


Figure 4.23: Trading strategy (elements of matrix Q that are greater than 0.5)

It was not tested whether this trading strategy works because we do not have a simulation of the trading process. We would need to send the orders ourselves based on our trading strategy. However, these orders could influence the market, so the orders (placed by other traders) coming after ours could be different from those that are in the data set.

Conclusion 5

In this thesis, we focused on modelling limit order books. We examined a data set with orders for futures contract for a German governmental bond.

In the first part of the thesis, it was described what a limit order book is. It consists of orders to buy or sell a certain asset. Each order usually contains information about the side, price, volume and time. A limit order book is the collection of all orders that are still waiting at a given time to be matched to some order on the opposite side. Some basic definitions regarding limit order books were introduced, such as the best buy and sell prices, mid-price, spread or imbalance index.

In Chapter 3, it was described how the algorithm for a limit order book construction works and how it processes three different types of requests: adding orders, deleting orders and modifying orders.

We also described how the mid-price and imbalance index that we get from the limit order book reconstruction can be modelled by a two-dimensional continuous-time Markov chain and how it can be transformed into a one-dimensional continuous-time Markov chain. From this model, we can estimate the probability of a future price change, provided that we know the current price change, the current imbalance index and the previous imbalance index. The trading strategy can be based on these probabilities.

In the last chapter, all results were shown. We described what kind of data set was available and how it was prepared for the limit order book reconstruction. Then it was shown what results we got from the limit order book reconstruction. It was shown how the limit order book looked at 10:00 and at the end of the day and how the best buy and sell price, mid-price and imbalance index evolved during the day. The best way to check whether the reconstruction was done correctly would be to compare it to the limit order book from Deutsche Börse AG, but no such information was available for this thesis.

Next, we tried to fit an exponential distribution to two sets of times: inter-arrival times (times of adding the orders) and the times the orders spent in the system (times between adding the order and deleting/matching it). It was not possible to fit it, so it was tried with power distribution. However, it could not be fitted as well. Therefore, it was tried to divide the inter-arrival times into smaller time intervals (1 hour, 15 minutes, and 10 minutes) and test the distribution inside these intervals. However, in most time intervals, the exponential distribution could not be fitted and power distribution was not fitted in any of them. This was not done with the second set of times because it would be difficult to find a way how to divide it into intervals. We also tried to test if some time intervals come from the same probability distribution. Each pair of intervals was tested, but in most of them, this hypothesis was rejected. Therefore, we did not look at the order flow as a Poisson process, but we modelled the price and imbalance change instead.

In the last part, a mid-price change and an imbalance index were modelled using a two-dimensional continuous-time Markov chain. From the data we got from the limit order book reconstruction, we estimated the probabilities of a future price change conditional on the current price change, the current imbalance index and the previous imbalance index. When the probability of a future price increase or decrease is greater than 0.5, respectively, we can buy or sell the instrument, respectively. This model depends on some choices of parameters, so it is possible that the results would change with a different choice. It was not tested how this trading strategy works because we did not have any simulation of the market and if we would place an order, it would affect the market and other traders' orders. Trying other models was beyond the scope of this thesis.

List of files A

This is the list of folders and files from the attached CD:

- `dp_2022_23_KUSOVA_Martina.pdf`: the text of this thesis
- `codes`: a folder that contains all MATLAB codes:
  - `lob_example.m`: an example of how we can use a bar chart to plot an LOB
  - `exp_dist_example.m`: plots of pdf and cdf of exponential distribution with a different choice of parameter λ
  - `power_dist_example.m`: plots of pdf and cdf of power distribution with a different choice of parameters a and k
  - `prob+`: a folder containing `PowerDistribution`:
    - `PowerDistribution.m`: customized MATLAB template to use a power distribution
  - `main_data_preparation.m`: loads `OrderAdd`, `OrderDelete`, `OrderModify` and `OrderSamePrio` and prepares them for LOB reconstruction (saves them as one table)
  - `data_preparation.m`: a function that is used in `main_data_preparation.m`, it deletes incorrect data, plots the prices and sizes in each file and puts `OrderAdd`, `OrderDelete`, `OrderModify` and `OrderSamePrio` into one table that is sorted by time
  - `main_reconstruction.m`: a reconstruction of LOB, it loads the table from data preparation, runs the algorithm of LOB reconstruction and saves the results into files
  - `algorithm.m`: a function that is used `main_reconstruction.m`, it contains the whole algorithm for LOB reconstruction, it calculates the best buy and sell prices, mid-price, spread and imbalance index in time, it saves the times when the orders were matched and it saves an LOB of a chosen level into a CSV file

  - main_plot_lob.m: plots an LOB at 10:00 (the data for that were saved manually by running the reconstruction only to a certain time) and at the end of the day, the best buy and sell price, mid-price, spread and imbalance index in time (the whole day and zoomed to an interval 10:00-10:05)
  - plot_lob.m: a function used for plotting in main_plot_lob.m
  - main_dist_test.m: tests the distribution of times
  - distribution_test.m: a function that is used in main_dist_test.m, it tests two sets of times for both exponential and power distribution: times of adding orders and times between adding and deleting/matching orders, then it tests the times of adding orders divided into smaller time intervals (1 hour, 15 minutes or 10 minutes) for both distributions and finally, it tests whether the times in some of these time intervals come from the same distribution (Kolmogorov-Smirnov test is used for all tests)
  - exp_distribution.m: a function that is used in distribution_test.m, it tests if times come from an exponential distribution using Kolmogorov-Smirnov test and plots their histogram and QQ plot
  - power_function.m: a function that is used in distribution_test.m, it tests if times come from power distribution using Kolmogorov-Smirnov test and plots their histogram and QQ plot
  - main_markov_chain.m: models mid-price and imbalance index using a Markov chain
  - markov_chain.m: a function that is used in main_markov_chain.m, it smooths the imbalance index, discretizes the smoothed imbalance index, calculates the price change, transforms the two-dimensional Markov chain to a one-dimensional, estimates the intensity matrix, calculates the probability of a future price change conditional on the current price change, the current imbalance index bin and the previous imbalance index bin and finds where this probability is greater than 0.5

- data-in: a folder that contains all input data (for a more detailed description of the data see section 4.1):

  - 2350_00_D_03_A_20191202.OrderAdd_FGBL_4128839.csv
  - 2350_00_D_03_A_20191202.OrderDelete_FGBL_4128839.csv
  - 2350_00_D_03_A_20191202.OrderModify_FGBL_4128839.csv
  - 2350_00_D_03_A_20191202.OrderModifySamePrio_FGBL_4128839.csv

- data-out: a folder that contains all data saved from data preparation and LOB reconstruction [1]:
  - add_FGBL_2019.mat: OrderAdd data after deleting the incorrect SecurityID
  - best_buy_FGBL_2019.mat: the best buy price in every time step
  - best_buy_FGBL_2019_10h.mat: the best buy price in every time step until 10:00
  - best_sell_FGBL_2019.mat: the best sell price in every time step
  - best_sell_FGBL_2019_10h.mat: the best sell price in every time step until 10:00
  - buy_FGBL_2019.mat: an LOB at the buy side at the end of the day (the first column contains prices and the second one sizes)
  - buy_FGBL_2019_10h.mat: an LOB at the buy side at 10:00 (the first column contains prices and the second one sizes)
  - delete_FGBL_2019.mat: OrderDelete data after deleting the incorrect SecurityID
  - FGBL_2019_lev5_orderbook.csv: a level 5 LOB in LOBSTER format
  - FGBL_2019_lev30_orderbook.csv: a level 30 LOB in LOBSTER format
  - I_FGBL_2019.mat: the imbalance index in every time step
  - I_FGBL_2019_10h.mat: the imbalance index in every time step until 10:00
  - mid_price_FGBL_2019.mat: the mid-price index in every time step
  - mid_price_FGBL_2019_10h.mat: the mid-price index in every time step until 10:00
  - sell_FGBL_2019.mat: an LOB at the sell side at the end of the day (the first column contains prices and the second one sizes)
  - sell_FGBL_2019_10h.mat: an LOB at the sell side at 10:00 (the first column contains prices and the second one sizes)
  - spread_FGBL_2019.mat: the spread in every time step
  - spread_FGBL_2019_10h.mat: the spread in every time step until 10:00
  - t_FGBL_2019.mat: the TrdRegTSTimeIn column from the table containing OrderAdd, OrderDelete, OrderModify and OrderModifySamePrio

[1] the files from 10:00 (file names ending with 10h) were saved by running only certain steps of the algorithm

- `t_FGBL_2019_10h.mat`: the `TrdRegTSTimeIn` column from the table containing `OrderAdd`, `OrderDelete`, `OrderModify` and `OrderModifySamePrio` until 10:00
- `tab_FGBL_2019.mat`: a table containing `OrderAdd`, `OrderDelete`, `OrderModify` and `OrderModifySamePrio` sorted by `TrdRegTSTimeIn` column
- `times_FGBL_2019.mat`: the times when the orders were matched (the first column contains IDs and the second one times when they were matched)

Bibliography

1. GOULD, Martin D. et al. *Limit Order Books*. 2013. Available from arXiv: 1012.0349 [q-fin.TR].

2. ABERGEL, Frédéric. *Limit order books*. Limit order books. 1st ed. Delhi: Cambridge University Press, 2016. Physics of society : econophysics and sociophysics. ISBN 1-316-87068-5.

3. NTAKARIS, Adamantios; MAGRIS, Martin; KANNIAINEN, Juho; GABBOUJ, Moncef; IOSIFIDIS, Alexandros. Benchmark Dataset for Mid-Price Prediction of Limit Order Book data. *CoRR*. 2017, vol. abs/1705.03233. Available from arXiv: 1705.03233.

4. RUBISOV, Anton. *Statistical Arbitrage using limit order book imbalance*. 2015. Available also from: https://tspace.library.utoronto.ca/bitstream/1807/70567/3/Rubisov_Anton_201511_MAS_thesis.pdf.

5. LOBSTER. *output.* [N.d.]. Available also from: https://lobsterdata.com/info/DataStructure.php. Accessed: 17. 3. 2023.

6. MÁLEK, Jiří. *Opce a futures [Options and futures]*. Druhé vydání. Oeconomica, 2003.

7. RADOVÁ, Jarmila; DVOŘÁK, Petr; MÁLEK, Jiří. *Finanční matematika pro každého [Financial mathematics for everyone]*. 8. rozšířené vydání. Grada Publishing a. s., 2013.

8. WOLFRAM RESEARCH. *PowerDistribution*. 2016. Available also from: https://reference.wolfram.com/language/ref/PowerDistribution.html. Accessed: 17.4.2023.

9. REIF, Jiří. *Metody matematické statistiky [Methods of mathematical statistics]*. 2. upr. vydání. Západočeská univerzita, Fakulta aplikovaných věd, 2004.

10. LILLIEFORS, Hubert W. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association* [online]. 1969, vol. 64, no. 325, pp. 387–389 [visited on 2023-04-11]. ISSN 01621459. Available from: http://www.jstor.org/stable/2283748.

11. LIKEŠ, Jiří; LAGA, Josef. *Základní statistické tabulky [Basic statistical tables].* První vydání. SNTL, 1978.
12. *Machine learning for Statistical Arbitrage II: Feature Engineering and Model Development*. [N.d.]. Available also from: https://www.mathworks.com/help/finance/machine-learning-for-statistical-arbitrage-ii-feature-engineering-model-development.html#mw_58c268cf-1cab-4489-9a15-84f0a80c14b7.
13. *Enhanced Order Book Interface*. Deutsche Börse Group, 2019. Available also from: https://www.xetra.com/resource/blob/1741884/e4a1e8329d20a0b2a052492956eb9beb/data/T7_EOBI_Manual_v.8.0.4.pdf.
14. *Epoch converter*. [N.d.]. Available also from: https://www.epochconverter.com/.
15. *Euro-Bund Futures*. [N.d.]. Available also from: https://www.eurex.com/ex-en/markets/int/fix/government-bonds/Euro-Bund-Futures-137298. Accessed 3. 5. 2023.

List of Figures

2.1 An example of LOB at time t 9

2.2 An example of orders at a certain price in LOB at time t 9

2.3 Exponential distribution with different values of parameter λ 13

2.4 Power distribution with different values of parameter k 14

2.5 Power distribution with different values of parameter a 14

4.1 The price and size of FGBL in 2. 12. 2019 in time in `OrderAdd` file . . . 26

4.2 The price and size of FGBL in 2. 12. 2019 in time in `OrderDelete` file . 27

4.3 The price and size of FGBL in 2. 12. 2019 in time in `OrderModify` file . 27

4.4 The price and size of FGBL in 2. 12. 2019 in time in `OrderModifySamePrio` file 28

4.5 The LOB at 10:00 28

4.6 The LOB at the end of the day 29

4.7 The best buy and the best sell price in time 30

4.8 The mid-price, the spread and the imbalance index in time 30

4.9 The best buy and the best sell price in time between 10:00 and 10:05 . 31

4.10 The mid-price, the spread and the imbalance index in time between 10:00 and 10:05 31

4.11 The best buy price, the best sell price and the mid-price in time between 10:00 and 10:05 32

4.12 QQ plot of the quantiles of the fitted exponential distribution vs. quantiles of the times of adding orders 33

4.13 The result of a test that times of add orders divided into 15 minutes intervals come from the same distribution (blue dots mean rejecting the null hypothesis and blank spaces mean rejecting) 36

4.14 P-values of a test that times of add orders divided into 15 minutes interval come from the same distribution 36

4.15 Histograms of two chosen intervals where the null hypothesis was not rejected 37

4.16 QQ plot of quantiles of data in intervals 0:45 - 1:00 and 2:00 - 2:15 where the null hypothesis was not rejected 37

4.17 Histograms of two chosen intervals where the null hypothesis was rejected 38
4.18 QQ plot of quantiles of data in intervals 0:30 - 0:45 and 0:45 - 1:00 where the null hypothesis was rejected 38
4.19 Mid-price and imbalance index in time intervals 10:00-16:00 and 10:00-10:05 39
4.20 Smoothed imbalance index in time intervals 10:00-16:00 and 10:00-10:05 40
4.21 Smoothed imbalance index, discretized imbalance index and mid-price change in time intervals 10:00-16:00 and 10:00-10:05 40
4.22 The probabilities of the future price change conditional on the current price change, the current imbalance index bin and the previous imbalance index bin 42
4.23 Trading strategy (elements of matrix Q that are greater than 0.5) . . . 42

List of Tables

2.1 LOBSTER format 10

2.2 The data from Figure 2.1 displayed in LOBSTER format with level 2 . 10

3.1 An example of encoding with 3 bins for imbalance index 20

4.1 A sample of order add data 24

4.2 Number of messages in the FGBL 2. 12. 2019 order files 25

4.3 Level 5 LOB at 10:00 28

4.4 Level 5 LOB at the end of the day 29

4.5 The proportion of time intervals where the null hypothesis that the times of add orders come from a fitted exponential distribution was not rejected 35

4.6 The proportion of the pairs of time intervals where the null hypothesis that the times of add orders come from the same distribution was not rejected 38

4.7 The probabilities of the future price change conditional on the current price change, the current imbalance index bin and the previous imbalance index bin 41

za jednotku času pro $x \in \langle x_0, x_0+h \rangle$

okamžitá změna veličiny y v čase $x = x_0$. Př. a) Nechť

vztahem $s = s(t)$. $s(5)$ t_0 t_1 $s(6)$ Pak $\frac{s(t_0+h)-s(t_0)}{h}$ je prů

hmotného bodu v čase t_0. b) Nechť vztah $q = q(t)$ u

čase t Pak $\frac{q(t_0+\Delta t)-q(t_0)}{\Delta t}$ je průměrná změna

náboje, tj. proud : $q'(t_0) = i(t_0)$. c) Hmotnost radioak

hmotnosti za jednotku času je

existuje. Pokud existuje jen

v bodě $A = [x_0, f(x_0)]$. Normála grafu funk

A a je kolmá na tečnu y, tj. $k_n = -\frac{1}{k_t}$, u rovn

a f je spojitá v x_0) $\frac{1}{\sqrt[3]{x}}$, je tečna rovno

y (resp. x).

f nabývá v bodě $x_0 \in D(f)$ lokálního minima (resp. maxim

-"- ostrého lokálního minima (resp. maxima) $\overset{def}{\Longleftrightarrow} \exists \delta > 0$