Studies in informational price formation, prediction markets, and trading

Peter Antony Bebbington

Department of Physics and Astronomy

University College London

A thesis submitted in partial satisfaction of the degree of

Doctor of Philosophy

date

Abstract

This thesis is a collection of three separate studies – but split into four chapters – which address the underlying issues in the nature and dynamics of markets. The studies investigate price-formation in the presence of noisy asymmetric information flow to a synthetic market, the statistical behaviour of in-play predictive markets and a reformulation of the Markowitz portfolio optimisation for financial market securities into the time-domain.

The first study looks to examine modern *in-play* gambling or predictive markets, in particular, horse racing markets. Since the advent of online sports gambling approximately 15 years ago large amounts of data have been collected for many different sporting events such as football, greyhound racing and cricket. In this study, the focus is on in-play horse racing markets where stylised statistical facts are presented and discussed. Price efficiency is analysed, and statistical arbitrage trading algorithms are developed to evaluate such efficiencies/inefficiencies. We develop a new model for testing the efficiencies of the initial implied odds quoted on the market. Exploring the efficiencies/inefficiencies found in the in-play markets we develop a martingale toy model and a statistical arbitrage trading model.

In the second study, we explore price-formation and the pioneering approach to financial asset pricing known as the *Brody-Hughston-Macrina* framework. The Brody-Hughston-Macrina information-based asset pricing framework is investigated in two parts; the first a development of a trading model and the other a generalisation of the information process that does not assume a linear rate-of-information flow. The trading model developed is a computational agent-based model that allows different configurations of agents to trade and hence create a synthetic market. The different configurations are explored by tracking the market price and times between adjacent trades with respect to changing certain model parameters, such as spread. The generalisation of the rate-of-information does not assume a linear function, as in the original Brody-Hughston-Macrina framework, but instead one that is non-linear in time. We estimate such a function from gambling market data and find it not to be a linear function. The non-linear Brody-Hughston-Macrina framework is fitted to winning horse odds signals.

The final study is motivated by recent advances in the spectral theory of auto-covariance matrices, and we are led to revisit a reformulation of Markowitz' mean-variance portfolio optimisation approach in the time domain. In its simplest incarnation, it applies to a single traded asset and allows to find an optimal trading strategy which, for a given return, is minimally exposed to market price fluctuations. The model is initially investigated for a range of synthetic price processes, taken to be either second order stationary, or to exhibit second order stationary increments. Attention is paid to consequences of estimating auto-covariance matrices from small finite samples, and auto-covariance matrix cleaning strategies to mitigate against these are investigated. Finally, we apply our framework to real world data.

Contents

List of Figures

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisors Prof. Ian J. Ford and Dr F.M.C. Witte for the continuous support of my PhD study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better advisors and mentors for my PhD study.

My sincere thanks also go to Dr Julius Bonart, Prof. Reimer Kühn, Dr Andrea Macrina, Prof. Tomaso Aste, and James Le Roux for many stimulating discussions and collaborations on these collections of studies. Also, I would like to thank Sam Priestley and Toby Aldous of [3] for their time, in-depth knowledge of gambling markets, and their abundant amounts of data – as which without this it would of be a much shorter thesis.

I would like to express my gratitude to the UCL Centre for Doctoral Training in Financial Computing & Analytics, and the people that manage the centre: Prof. Philip Treleaven and Mrs Yonita Carter for selecting me for the programme and their invaluable guidance. And not forgetting the fund body EPSRC that without their financial support this would not have been possible.

Last but not the least, I would like to thank my family: my mother especially, Caroline Bebbington, for giving birth to me in the first place and supporting me throughout my life.

Chapter 1

Introduction

This thesis is a collection of studies that have the focus of markets and price formation. The markets explored are financial markets and in-play horse racing gambling markets.

The main core of the work is broken down into four chapters with the following titles: Chapter 2, *Statistical Analysis of Horse Gambling In-play*; Chapter 3, *Simulating In-Play Horse Odds and A Novel Trading Algorithm*; Chapter 4, *Information Based Finance and Trading*; and Chapter 5, *Optimal Trading Strategies – a Time Series Approach*. Each of these studies involves different approaches to investigate the notion of price and how it may be perceived differently by independent agents, making it a social construct of markets. These studies explore two different of types markets: gambling and financial. Markets are exchanges where collections of interacting agents buy and sell assets, and through this medium, assigning the assets with a subjective value. The work focuses solely on this concept otherwise known as *market price* and uses a combination of statistics and stochastic dynamics to understand its features and behaviour. We introduce these studies individually by giving the reader a brief summary of the content of each of the chapters.

Chapter 2 *Statistical Analysis of Horse Gambling In-play Markets* initially gives a background analysis of the gambling industry in Table 2.1 but then turns the focus to the exchange markets known as *in-play horse racing markets*. The same section also discusses the mathematics of a traded asset in this market referred to as the *odds*. We discuss that odds are similar – but not equivalent – to financial *market price*; as the odds quoted have a value that can be extracted through hedging techniques, to generate gains. The dataset used in this study consists of 12736 races that took place across the British Isles during the period starting 31-12-11 and finishing 31-12-12, this data was provided by [3]. The activity that coincides with the time in the race after it has begun is known as the in-play markets, and this temporal part of the market is paramount to this chapter's analysis.

The data set is explicitly described in Sec. 2.2 where a detailed discussion of the variables is given in Table 2.1. In Sec. 2.2.1 we illustrate the typical signal dynamics found in horse racing in-play markets to give the reader an insight into how the market evolves in time.

The sample statistics are estimated in Sec. 2.3 in Table 2.2, we also investigate the initial odds quoted on the market and compare this with the true odds of this horse winning the race. Using statistical dispersion measures, such as the Gini coefficient, we show in sec 2.3.1 that on average the uncertainty of the in-play race odds signals decreases in time.

The last section, Sec. 2.4, published in the paper [4], we find – a somewhat surprising result – that the initial odds quoted on the in-play market have a distribution that is similar to the segments of a randomly divided interval.

Chapter 3 *Simulating In-Play Horse Odds and A Novel Trading Algorithm* continues in the same spirit as the previous chapter but here we look at modelling the odds of the in-play horse racing signals.

Sec. 3.1 gives a review of the modern economic theory known as the *Efficient Market Hypothesis* (EMH), which in short assesses whether a market price is (or can be) unduly manipulated by market agent/agents that possess information that other agents do not hold. We explain that a market price that is considered to be efficient is known as a *martingale* and such process are used to test if EMH holds true. If it assumed that EMH holds true, then one can propose a model using synthetic stochastic processes to generate a computationally based horse race. This computationally based horse race is performed in Sec. 3.2 and we demonstrate that using Itô processes as a model of a horse's performance and a Monte Carlo method one can estimate the odds signals for a particular horse winning.

Sec. 3.3 gives a review of the pair trading algorithm from a financial market point-of-view. It discusses a linear pricing model and how these models can be used to determine the necessary quantity of equity to hold to be market risk neutral. The section then moves on to discuss methods of detecting pairs comparing the statistics of signals to those of random walks. Once all the theory of creating a pairs trading algorithm is in place, in Sec. 3.3, we turn our attention to the applicability of the model. This pairs trading algorithm leads us to apply the model to the in-play horse racing data described in Chapter 2. This application is discussed in Sec. 3.4 where we find the algorithm works remarkably well and generates high rates-of-return over many races.

Chapter 4 *Information Based Finance and Trading* reviews and develops a mathematical framework from the field of financial mathematics, namely *credit-risk*. We delve into a model established just over a decade ago known as the *Brody-Hughston-Macrina* (BHM) approach or information-based asset pricing.

The BHM approach is first reviewed in Secs. 4.1 and 4.2 where we have presented the background of literature that led to the initial development of the BHM framework and how it is used to price assets. The BHM model like all stochastic finance models is constructed from probability theory. These models are used for the pricing of financial instruments, and such prices are assumed to exist only if a risk-neutral probability density exists. The essence of what the BHM model tries to achieve is to treat market prices as the perception of market participants, and this perception is contrived as a filtration process known as the *information process*. In probability theory, information is formulated with filtrations, and commonplace to assume that these processes are adapted to an underlying stochastic price process; for example, a geometric-Brownian motion is used in the Black-Scholes-Merton model. The BHM framework assumes that the filtrations are adapted to a class of Markov processes which consist of two additive components: signal and noise. The signal component is theorised to be a random variable which represents a final outcome, or a cash-flow, that is partially revealed linearly in time. The noise component is a Brownian bridge process which is pinned to zero at both ends, and this process provides noise variance that is linearly increasing in time up to the half way point and then linearly decreasing in time to the end. Assuming a risk-neutral pricing kernel exists and conditioning this on the theorised information process generates a price which dynamically evolves as an emergent phenomenon, which will be discussed in more detail in Sec. 4.1 and 4.2.

We develop the BHM framework in two ways: firstly by creating a mechanism for trading to occur between agents that are pricing the same asset in Sec. 4.3 and secondly a refinement in the BHM model where the assumption that information is revealed linearly in time is relaxed Sec. 4.6. The toy model developed in Sec. 4.3 establishes a synthetic market where the market price is aggregated from exchanges between agents that price according to the BHM pricing model. Once a trade occurs in our synthetic market, the prices of the two trading agents are updated accordingly, with a shift to the average price of the two agents at the time of trading. Different market configurations are considered with three categories of market participants: Market Maker, Informed Traders, and Noise Traders who have their individual trading rules and parameterisation. We also introduce inventory control in Sec. 4.3.6 which is used as a means to control the overall frequency of agents' trading, especially market makers with small spreads. The results for these synthetic markets are presented in Sec. 4.4 with plots of distributions for the increments of market price and first passage time.

Sec. 4.5 discusses and highlights limiting properties of the BHM framework, such as the linear revelation of information. We suggest a refinement, or generalisation, in 4.6 where we take the linear assumption, in the BHM model, and replace it with a non-linear function. This relaxation of this assumption, linearly developing perception of the final outcome, allows one to derive a new price dynamic which is more flexible in how information is affirmed in the market. The final section, Sec. 4.7, takes the original BHM model and the non-linear BHM model and explores fitting these to in-play horse racing price signals using the same data [3].

The final chapter of this thesis: Chapter 5 *Optimal Trading Strategies a –*

Time Series Approach presents a mean-variance optimisation model – similar to Markowitz' portfolio optimisation theory [5] – applied to a single time period to find an optimal trading strategy, presented in the paper [6]. An alternative approach to when seeking an optimal trading strategy for capital allocation as compared the to the popular dynamic programming approach that requires solving a Hamilton-Jacobi-Bellman or Bellman equation [7–14]. The Markowitz portfolio optimisation theory has a rich history in economic research and industrial practice [15–19]. One of the main reasons for its popularity is clearly its conceptual simplicity, which helps in building an intuition about the nature of risk and its relation to an investment's return.

A review of time series models is given in Sec. 5.1 which highlights concepts such as weak stationarity, auto-correlations and stationary time series models. A brief description of the Markowitz approach to portfolio optimisation is given in Sec. 5.2, and how it is translated into the time domain. We proceed to apply this time translated model in Sec. 5.3 where it is first implemented with synthetic price signals that are stationary and stationary in their increment.

The results of this numerical investigation for synthetic processes provides an insight into the influence of sampling noise on optimal strategies and risk-return profiles. In Sec. 5.4 we look at optimal trading strategies for empirical data, using the S&P500 index as an example and investigate the effect of auto-covariance matrix cleaning on risk-return profiles, based on comparing auto-covariance spectra for the S&P500 and expected spectra for a process with uncorrelated increments. We finally apply a cleaning methodology to the estimation of empirical auto-covariance in Sec. 5.4.2, which improves the estimation of the trading strategies.

Chapter 2

Statistical Analysis of Horse Gambling In-play Markets

This chapter is split into four parts in which we briefly explore the gambling industry and take a more thorough look at the statistics of in-play horse gambling markets. The first part will give a review and background to the gambling industry and in particular modern online *peer-to-peer* gambling markets 2.1, such as Betfair [20]. The second part we present the reader with a set of actual signals from a typical in-play horse race and outline the details of the horse racing data set this is available to this study. The penultimate part gives a detailed study of the statistics that are generated by in-play horse racing markets. The statistics are estimated from the variables discussed in Sec. 2.2. The final part investigates the behaviour in the order statistics, see [21], of the initial odds and draws a comparison between the expectation of the random division of an interval.

2.1 Online Gambling Markets Review

Over the last two decades the industry of *peer-to-peer* gambling has changed from a market that was almost completely paper traded to one that is almost completely digitally traded [22]. The global growth of the Internet gambling market has increased every year since 2000 [22], and mobile markets alone have been estimated to reach 45% of entire activity by 2018 [23]. This digital paradigm shift has led to vast amounts of market data becoming available, opening the door for large-scale analysis of historical trading behaviour in gambling markets. This trading behaviour can be analysed by constructing models to help interpret how agents react to information. Trading behaviour patterns can be thought of as the market agents' perceptions and the changes in those perceptions to market information. These changes in market information may be observed as a linear or non-linear change in the odds that are quoted on the market, similar to what is seen in financial markets.

Gambling markets consist of many different games for which odds are exchanged, for example, football, tennis, greyhound racing and horse racing. The data that is used in this study are solely for horse racing and are described in detail throughout this chapter, particularly in Sec. 2.2. It is worth noting that not all markets behave in the same way. For example, horse racing and tennis will have very different price dynamics, but horse racing and greyhound racing would be very similar. Why such differences emerge through price is because different features and rules exist in various competitions and games and how the market processes the information flow leads to different price dynamics. Consider the tennis match and horse race one of the main difference is the time duration as tennis matches are usually much longer. Another difference is how the competitions are won, as in a race the competitor is competing for the pole position, and in tennis the player has to win a certain number of sets.

It is understood that gambling can be addictive [24] and as it has become more accessible [25] the risk factors leading to social issues increase. The moral point made to the reader is that this study does not aim to encourage gambling and we are only interested in the dynamics and statistical behaviour of the odds.

As mentioned previously the data used in this study comes from a horse racing market and was taken from exchange known as Betfair [20]. Since the internet boom gambling markets have become modernised, leading to markets where odds are traded electronically. These modernised markets are referred to as *peer-to-peer* online markets. The modernisation has allowed market participants to have the freedom to buy odds but also to sell odds, which in gambling parlance is backing and laying respectively. Since agents can buy and sell odds (or chance), this means any agent in the market can act as a bookie by laying odds on the market. A market participant that buys and sells odds is a *market maker*, and this is a trading strategy that bears inventory risk but is rewarded by the size of the spread. The markets maker is rewarded with the spread because they broker the deal, selling at the asking price and buying at the bidding price securing the difference between the two. This technological change has massively transformed the structure of the gambling industry all over Europe [26] as compared to the market in the pre-internet age. It has been noted that these changes in technology have brought about radical new business models and trading strategies based on the information efficiency in the peer-to-peer betting exchanges [27]. A review of the global gambling markets can be found in Chapter 2 pg. 7-25 in [22] which gives a breakdown of the different betting markets; for example, online gambling, casino games and gaming machines. The same chapter in [22] also provides estimates of regional internet market share, and within Europe, the UK is expected to be the largest. As a final note, we would like to point out that this research has no ties with Betfair – or another exchange – and has been conducted independently to any other private enterprises.

2.1.1 Investing and Gambling

For centuries mathematicians, physicists and economists have been exploring systems that are considered to be random. This interest started in the seventeenth century where the French mathematicians Pierre de Fermat and Blaise Pascal developed the notation of expectation. Even the famous economist John Maynard Keynes had this to say about gambling and its relationship to investing

> *"It is generally agreed that casinos should, in the public interest, be inaccessible and expensive. And perhaps the same is true of stock exchanges." - John Maynard Keynes [28]*

In the past there has been a clear dichotomy that gambling was strictly regarded as an entertainment industry and investing as a business activity but in modern gambling this split is not so clearly defined [2, 29]. There are reasons for the apparent cloudiness in definitions, but it must be noted that these two areas are clearly different phenomena though they do share some similar behaviours. These similar behaviours come from the fact that both investing and gambling involve the interactions between people and markets. A simplified definition of the dichotomy of investing and gambling is as follows:

- Investing is the process of using capital to purchase a financial instrument that potentially can yield a return.
- Gambling is the process of using money as a stake in the outcome of an event which if correct yields a return.

Such a set of crude definitions can confuse one into thinking that investing and gambling are the same, but one must think of the big picture, as this is where the two clearly diverge. Although investment and gambling play very different roles in the economy, that does not necessarily impinge on the nature of the "decision-making under uncertainty" aspect of these two phenomena. Investing is the driving mechanism of *capitalism*, this is how capital circulates around the global markets, and capitalism is society's means to prosperity.[*] Prosperity and higher standards of living are achieved through investment by aiding the growth of an economy which in turn creates more jobs. Professionals that work within the financial markets serve the purpose of making investments. Besides, investments have an associated risk, some of which can be hedged using particular trading strategies [32]. A dynamic economy has frequent investment, and this behaviour increases market liquidity and price efficiency, with the secondary characteristic of more transparency within markets as greater amounts of information regarding investments is exchanged and published through news outlets. Since the advent of the "financialisation" of the economy, investment is regularly the acquisition of certain types of financial products; this may have only very slim relationship with more traditional "investment", such as investment in infrastructure projects, technological development or productive capital such as machinery.

[*] This point is obviously and deliberately oversimplified as the role of capitalism and prosperity in society is not as clear cut as this and is a very complex issue which can be seen in [30, 31].

Gambling, on the other hand, does not have this direct positive *net* effect upon the global economy. One could argue that its rise in popularity through online peer-to-peer gambling sites has increased employment within a *local* economy, but it does bring with it the negative side effects such as addiction [22, 24, 33]. A bet is in many ways just a contract between two parties where, in exchange for a fee, one party promises the other a payout depending on the realisation of a future "state of the world".[†] Placing capital on an event in a gambling game doesn't affect the final results[*]; consider a horse race if one places more backs on the leading horse this obviously does not affect the performance of the horse. Investments on the other hand, generally speaking, can have a positive effect on the performance of a company. Consider the example of a factory with a surplus demand but not the additional capital to buy more machinery to help meet this demand. With the allocation of capital from an investor the factory can purchase the equipment and hire the extra people needed to meet this additional demand, thus making the factory more productive and new employees more prosperous.[§]

To summarise the main points of the roles of gambling and investment in society and their inherent similarities and differences between them:

1. Both are the pursuit of return via a risky decision regarding a particular outcome of a random event in the future.

2. In general more research goes into investing, such as technical analysis, which helps to ensure that the risk of losing capital is more in favour of the investor.[‡]

3. Investing potentially has a positive net social and economic effect whereas gambling has less. [**]

[†] Similar contracts are also signed in the insurance business.

[*] One can argue that in a poker game a player can win because of the bluffing strategy and the outcome of who wins changes but this does not alter the distribution of the cards.

[§] This is the classical version of investment.

[‡] Put another way gambling is a risk-seeking strategy whereas investing is a risk-averse one.

[**] But note that the gambling industry is quite generally part of the services sector of the economy. All transaction fees paid add to GDP, creating jobs and generating tax revenue.

4. Investing is the ongoing and long-term flow of capital, whereas gambling is short term and has a terminal (stopping) time when all the processes end.

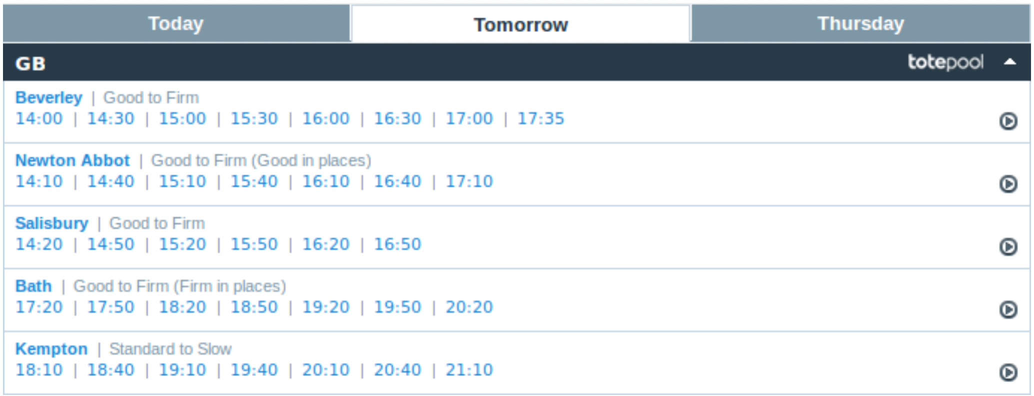
5. Investment is a process of business and commerce, and hence the global economy and gambling is a market share within the entertainment sector.

2.1.2 Horse Racing Markets

Horse racing markets will be the focus of this study, but before this analysis is done, it is important to know how this particular market operates. The biggest gambling market operating in Europe is Betfair [20] and the horse racing exchange is found at this URL `https://www.betfair.com/exchange/horse-racing`. The site displays the next three days of races from all across the world. The location of the races from which the market data has been collected is from tracks located across Great Britain, Northern Ireland and the Republic of Ireland. Fig. 2.1 shows a screen shot for a typical day of a Betfair horse racing exchange within race tracks found in Great Britain, the interface displays the different race track names and race start times. When selecting a race time the user is taken to the race order-book, an example of which is shown in Fig. 2.1. The order-book is a set of market odds (also referred to as prices) which are offered by agents. The order-book acts as an auction where agents can back (buy) or lay (sell) odds for each competing horse (*selection*) in the race. The order-book is split into two halves with three levels of back prices and three levels of lay prices for each selection. The most competitive odds are found in the middle of the order-book with blue being the best back and pink the best lay, and the distance between them is the spread defined in Eq. (2.14). Each of the odds defined in the order-book has a volume associated with them and are in the units £. The volume informs the market how much can be taken at a particular level of odds.

The relationship between volume and odds can be observed by considering Fig. 2.2 and Fig. 2.3 and the top selection Storm Melody (one of the favourites in the order-book). If an agent wishes to buy the odds for this selection the best bid lies at the value 4.9* with a total volume of £43: if said agent places a market

*This is a known as decimal odds and is explained in more detail in Sec. 2.1.3.



Today | Tomorrow | Thursday

GB — totepool

Beverley | Good to Firm
14:00 | 14:30 | 15:00 | 15:30 | 16:00 | 16:30 | 17:00 | 17:35

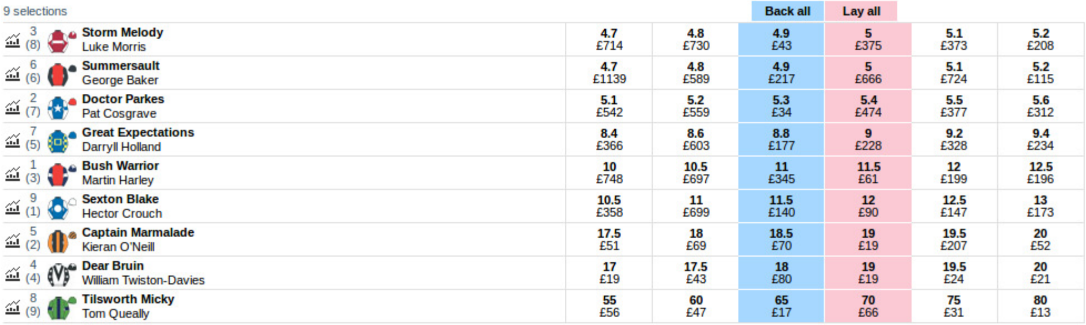
Newton Abbot | Good to Firm (Good in places)
14:10 | 14:40 | 15:10 | 15:40 | 16:10 | 16:40 | 17:10

Salisbury | Good to Firm
14:20 | 14:50 | 15:20 | 15:50 | 16:20 | 16:50

Bath | Good to Firm (Firm in places)
17:20 | 17:50 | 18:20 | 18:50 | 19:20 | 19:50 | 20:20

Kempton | Standard to Slow
18:10 | 18:40 | 19:10 | 19:40 | 20:10 | 20:40 | 21:10

Figure 2.1: The interface of a typical day of horse races across Great Britain. The interface has a tab that allows the user to select from the three days Today, Tomorrow and the day after tomorrow. On a particular day, there are track location names, for example, the Tomorrow tab that is highlighted gives the choices to select from Beverley, Newton Abbot, Salisbury, Bath and Kempton. Each track name has a list of times below them which are the start times of each race at that track location for the day and selecting one of them takes the user to the order-book an example of which is shown in Fig. 2.2.



9 selections			Back all	Lay all		
3 (8) Storm Melody Luke Morris	4.7 £714	4.8 £730	4.9 £43	5 £375	5.1 £373	5.2 £208
6 (6) Summersault George Baker	4.7 £1139	4.8 £589	4.9 £217	5 £666	5.1 £724	5.2 £115
2 (7) Doctor Parkes Pat Cosgrave	5.1 £542	5.2 £559	5.3 £34	5.4 £474	5.5 £377	5.6 £312
7 (5) Great Expectations Darryll Holland	8.4 £366	8.6 £603	8.8 £177	9 £228	9.2 £328	9.4 £234
1 (3) Bush Warrior Martin Harley	10 £748	10.5 £697	11 £345	11.5 £61	12 £199	12.5 £196
9 (1) Sexton Blake Hector Crouch	10.5 £358	11 £699	11.5 £140	12 £90	12.5 £147	13 £173
5 (2) Captain Marmalade Kieran O'Neill	17.5 £51	18 £69	18.5 £70	19 £19	19.5 £207	20 £52
4 (4) Dear Bruin William Twiston-Davies	17 £19	17.5 £43	18 £80	19 £19	19.5 £24	20 £21
8 (9) Tilsworth Micky Tom Queally	55 £56	60 £47	65 £17	70 £66	75 £31	80 £13

Figure 2.2: An example of a Betfair order-book for a horse race which is the Beverley 14:00 race from Fig. 2.1. This race had nine selections, or runners, with the top selection being the current favourite. The left side starting from the blue boxes are the back odds on offer with the best back odds in the blue boxes. The right side starting from the pink are the lay odds with the best lay odds in the pink boxes. The odds are in a decimal form which is the reciprocal of the implied probabilities (see Sec. 2.1.3) and are located in bold at the top of the boxes. The volume is located below the odds and is in terms of pounds sterling £ and this is the total amount available on the selection at that price.

order for £43 at 4.9, then that box will be empty/exhausted. The empty box can lead to two possible outcomes: a *limit-order* comes in at the original odds 4.9 with a volume £B and the spread remains unchanged at a value of one tick or 0.1, or a limit-order comes in at 4.6 with a volume £C and the spread increases by two ticks or 0.2. These two possible situations are shown graphically in Fig. 2.3. There is also a third possibility which is the box remains empty, that is the odds and volume do not shift along to occupy the three best backs, but this depends on the liquidity/market activity for the selection and also the market's ability to convert the information of the selection's performance into a market price[†]. If liquidity is high on this selection and it is performing well the most likely outcome is the top result in Fig. 2.3. Notice that if, in Fig. 2.3, the bottom situation occurs, the decimal odds decrease to 4.8 which means the subjective likelihood of this selection winning increases.

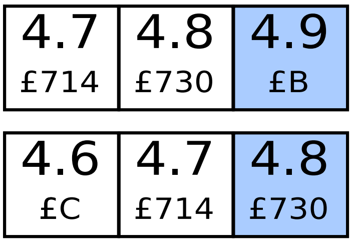


Figure 2.3: The backing side of the order book for the selection Stormy Melody in Fig. 2.2 after the best back price is completely exhausted. The top and bottom images show the two possible outcomes where the volumes B and C are positive integers.

2.1.3 Odds

Odds are the assets that are exchanged in gambling market and can be considered binary options [35]. In this section, we will look at how online gambling markets define implied odds and subjective probabilities. Consider a horse race that is indexed as $i \in \{1, 2, \dots, N_h\}$ where N_h is the total number of runners in the race.

[†] Here it is assumed that odds in race track markets exhibit price efficiency otherwise known as the efficient market hypothesis [29, 34] and Sec. 3.1.

The *belief* that a particular horse will win the race is denoted as

$$P_t^{(i)} = \mathbb{P}\left(X_T^{(i)} = 1\right) \text{ such that } \sum_{i=1}^{N_h} P_t^{(i)} = 1^{*} \tag{2.1}$$

and is an implied probability. One way to calculate this quantity is using the *parimutuel betting* system. This framework takes all the bet volumes denoted as $V_1, V_2, \ldots, V_{N_h}$ and sums them up: $V_{\text{Total}} = \sum_i^{N_h} V_i$. Normally the market maker or bookie will take a fraction of the total of bets V_{Total}: this fraction is denoted as $c \in (0,1)$. Hence, under a parimutuel betting framework, the implied parimutuel probability is

$$P_t^{(i)} = \frac{V_i}{(1-c)V_{\text{Total}}} \tag{2.2}$$

where $P_t^{(i)} \in (0,1)$ and t is the time dependence. In the Betfair in-play market, participants would not use a parimutuel system but instead set the odds using a double-auction mechanism with market makers [36,37] as described in Sec. 2.1.2. In such a market the collective behaviour of the agents in the market set the back and lay prices in a two-way auction. This mechanism can also be observed by the fact *over-rounds* are so prevalent in the in-play markets, where the over-round is the excess value of the total implied probabilities of the selections that do not converge to unity.

If the implied probabilities sum as $\sum_{i=1}^{N_h} P_t^{(i)} = 1$ the race is a fair game but if for example $\sum_{i=1}^{N_h} P_t^{(i)} > 1$ then one has a *dutch-book* [27]. The dutch-book in horse racing is define numerically by the *over-round*, which if greater than unity [38] guarantees the bookies (sellers of odds) a profit no matter the outcome of the race. Generally speaking, gambling markets never use implied probabilities and for historical reasons work with either *decimal* odds or *fractional* odds. Betfair uses the *reciprocal* or *decimal* odds which are defined as

$$O_t^{(i)} = \frac{1}{P_t^{(i)}} \tag{2.3}$$

[*] This is not strictly true in real markets and the discrepancy is observed as the market *over-round*, see Fig. 2.4 top left plot.

where $O_t^{(i)} \in (1, \infty)$ are the odds seem in the order book shown in Fig. 2.2. If one has the fractional odds $a_t^{(i)} : b_t^{(i)}$ then the implied win and lose probabilities respectively are

$$P_t^{(i)} = \frac{a_t^{(i)}}{a_t^{(i)}+b_t^{(i)}} \text{ and } Q_t^{(i)} = \frac{b_t^{(i)}}{a_t^{(i)}+b_t^{(i)}} \tag{2.4}$$

where $Q_t^{(i)} + P_t^{(i)} = 1$ for each i. The odds of winning and losing are therefore the following ratios

$$\begin{aligned} a_t^{(i)}/b_t^{(i)} &= P_t^{(i)}/Q_t^{(i)} \triangleq W_t^{(i)} \\ b_t^{(i)}/a_t^{(i)} &= Q_t^{(i)}/P_t^{(i)} \triangleq L_t^{(i)} \end{aligned} \tag{2.5}$$

where one has defined the odds in favour as $W_t^{(i)}$ and odds against as $L_t^{(i)}$, such that $W_t^{(i)} L_t^{(i)} = 1$ and $\prod_{i=1}^{N_h} W_t^{(i)} L_t^{(i)} = 1$. There are now three scenarios for a particular horse: $W_t^{(i)} > L_t^{(i)} \Leftrightarrow P_t^{(i)} > Q_t^{(i)}$ (higher implied chance of winning), $W_t^{(i)} < L_t^{(i)} \Leftrightarrow P_t^{(i)} < Q_t^{(i)}$ (higher implied chance of losing) and $W_t^{(i)} = L_t^{(i)}$ (even implied chance of winning and losing). As a rule of thumb this is represented as

$$\begin{cases} 1 : W_t^{(i)} & \text{if } W_t^{(i)} > L_t^{(i)} \Leftrightarrow W_t^{(i)} > 1 \Leftrightarrow L_t^{(i)} < 1 \\ 1 : 1 & \text{if } W_t^{(i)} = L_t^{(i)} \\ L_t^{(i)} : 1 & \text{if } W_t^{(i)} < L_t^{(i)} \Leftrightarrow W_t^{(i)} < 1 \Leftrightarrow L_t^{(i)} > 1. \end{cases} \tag{2.6}$$

To explain this further, consider the example of a two horse race where $i = 1, 2$ and t is fixed:

1. If horse $i = 1$ has odds $1 : 4 \Leftrightarrow W_t^{(1)} = 4$ and $L_t^{(1)} = 1/4$ this translates to the implied probabilities $P_t^{(1)} = 0.8$ and $Q_t^{(1)} = 0.2$.

2. The previous statement implies that horse $i = 2$ has odds $4 : 1 \Leftrightarrow W_t^{(2)} = 1/4$ and $L_t^{(2)} = 4$ this translates to the implied probabilities $P_t^{(2)} = 0.2$ and $Q_t^{(2)} = 0.8$.

Decimal odds as used by Betfair the reciprocal implied probabilities are set by the market as in Eq. (2.3) and seen in Fig. 2.3. One of the practical advantages of decimal odds are they have floating point precision and hence require less computer memory to store than the converted implied probabilities which have double point precision. Decimal odds used on Betfair fix the smallest possible tick size to 0.01. One can work out the return on investment if horse i wins as following the percentage

$$R_t^{(i)} = (O_t^{(i)} - 1) \times 100\% \tag{2.7}$$

where the stake has been removed as this is returned if the selection wins.

2.1.4 Hedging and Arbitrage

This is the process that ensures that, in any outcome, a gambler is guaranteed to break even or make a profit from an trading strategy. This can be achieved by hedging with a simple strategy where one backs then lays on a selection. This strategy can put a gambler in a hedged position, also known as an arbitrage, meaning that they are guaranteed to profit no matter what the outcome. This hedge is considered to be successful if for any movement of the odds the bet is still profitable, hence the strategy is directionless with respect to the odds movement. The movement of the odds dictate the order by which the hedging transactions are executed, that is:

- If the odds are **Decreasing** → **Back** first then **Lay** second.
- If the odds are **Increasing** → **Lay** first then **Back** second.

Two common strategies are used in gambling markets to hedge one's stakes. The first one is known as "*Green Up*": this method is known as this because in a gambling exchange interface, such as Betfair, the price in the user's portfolio turns green to indicate an arbitrage/hedge position. The green up method requires two or more stakes to be performed in the following order; back or lay on the same selection in equal amounts, ensuring the back odds are larger than the lay odds and vice versa for the lay position first. Dividing the potential profit if the horse wins by the available lay odds you will see, if the orders have moved in your favour, all the prices on the exchange screen turn green. The method is outlined in the following steps:

(1) Place a back in a down trending market, or a lay in a up trending market.

(2) If the odds move in your favour, lay the same amount placed on the back in the down trending market, or back the same amount placed on the lay in the up trending market.

(3) This is equivalent to a risk-free bet meaning one is in a situation where one can only win (Arbitrage) or break even (Hedged).

(4) Dividing the profit of this risk free bet by the current lay odds will give the "Green up" screen.

To see why this is an arbitrage or a hedge, consider a horse in a given race which has the odds 9:1 and back this with £100. If the odds are in a down trend, say the odds drop to 8:1, lay £100 on the same selection. The two possible outcomes of the strategy and the selection will be:

- Winning horse ⇒ £100×9 and an exposure to the lay (liability) of £100×8 which means a profit of £100.
- Losing horse ⇒ £100 was gained in the lay and a loss of £100 in the back, hence break even.

This demonstrates a strategy with a profit which is floored at zero and ceiling of £100, hence such a strategy will not green up, but one can guarantee a profit by changing the volumes placed. A strategy that is completely hedged against losses but one that is guarantee a profit, a green up, is one where the gambler divides the back of £100 by the current lay odds 8:1, giving the lay stake £100/9.0=£11.11. Using this stake gives the two possible implications of the strategy, which are:

- Winning horse ⇒ £100×9 the liability £100×8 and £11.11×8 a guaranteed profit of £11.11.
- Losing horse ⇒ £100+£11.11 was gained in the lay and a loss of £100 in the back hence a profit £11.11.

This is an example of a successful green up as one would be able to sell back the original bet for £11.11 as this particular configuration of bets has given an arbitrage.

2.1.5 Gains

Deploying any sort of strategy one needs to keep track of returns and this section looks at how one can do this. A return is denoted as $R_t \in \mathbb{R}$ where t is a discrete time step dependence. A return in a prediction market[*] can be either a gain $R_t > 0$, a loss $R_t < 0$ or no return $R_t = 0$. Consider the following, if one backs at the time step t and lays at subsequent time step $t+1$[†] and the time index is $t \in \{1, 2, \dots, T-1\}$ where T is the final time step. This sequence is equivalent to the following sequence of trades $\{back_1, lay_2, back_3, lay_4, ...\}$. A few assumptions have been made here, first that successive *back* and *lay* positions can be held to the end of the event but in reality this may not always be the case as there could be liquidity restrictions; and secondly, one can only take binary positions which is defined as the following strategy function $\pi_t \in \{1, -1\}$ where $\pi_t = 1$ is a back position and $\pi_t = -1$ is a lay position[‡]. The next variable that needs to be defined is the volume or stake that is placed at each position which is denoted as $b_t > 0$, it is assumed that b_t is a constant and is small enough that the position is always matched[§]. The two scenarios can occur which are represented as the following matrices

$$\begin{array}{c} \\ X_T^{(i)}{=}1 \\ X_T^{(i)}{=}0 \end{array} \begin{array}{c} \begin{array}{cc} \text{Back} & \text{Lay} \end{array} \\ \left[\begin{array}{cc} O_t^{(i)} & -O_{t+1}^{(i)} \\ -1 & 1 \end{array}\right] \end{array} \text{ or } \begin{array}{c} \begin{array}{cc} \text{Lay} & \text{Back} \end{array} \\ \left[\begin{array}{cc} -O_t^{(i)} & O_{t+1}^{(i)} \\ 1 & -1 \end{array}\right] \end{array} \tag{2.8}$$

where $X_T^{(i)}$ is the win variable defined as

$$X_T^{(i)} = \begin{cases} 1 & \text{if } i \text{ wins} \\ 0 & \text{if } i \text{ loses,} \end{cases} \tag{2.9}$$

$O_t^{(i)}$ are the decimal odds Eq. (2.3) and $i \in \{1, 2, ..., N_h\}$ is the horse index. From a financial point of view these two matrices are equivalent to going *long* and *short* to a volume of one pound sterling. The change in the decimal odds is denoted as $\Delta O_t^{(i)} = O_{t+1}^{(i)} - O_t^{(i)}$ and it is trivial to see that $\Delta O_t^{(i)} = \Delta R_t^{(i)}$ from Eq. (2.7).

[*] Prediction markets are just another name for gambling markets.

[†] As a note it does not have to be the very next time step but any succeeding time step.

[‡] One could also define a third state $\pi_t = 0$ which is no position.

[§] The terminology of matched means a back or lay limit-order is exchanged in the order-book.

Applying the long strategy, which is the left matrix in Eq. (2.8), for a single time step from $t = 1$ to $t = 2$ the odds at each time step are thus $O_1^{(i)}$ and $O_2^{(i)}$. If one stakes b at both time steps for the losing horses the net gain is zero, but the gain if the horse wins will be $g_1 = b\left(O_2^{(i^*)} - O_1^{(i^*)}\right) = b\Delta O_1^{(i^*)} = b\Delta R_1^{(i^*)}$*. Therefore the gain for each long position or short position is only acquired from the winning horse signal which is determined at time $t = T$ and thus the returns are

$$R_t^{(i^*)} = b\left(\Delta O_t^{(i^*)}\right) \quad \text{or} \quad R_t^{(i^*)} = -b\left(\Delta O_t^{(i^*)}\right) \tag{2.10}$$

where the left expression is the gain on a long position and the right expression is the gain on a short position. The expression will give a positive gain in a long position when $O_t^{(i)} < O_{t+1}^{(i)}$ (the decimal odds move up in price) and if a short position $O_t^{(i)} > O_{t+1}^{(i)}$ (the decimal odds move down in price). Thus the accumulative gains on each strategy is the sum of long (or short) positions

$$R_T^{(i^*)} = \sum_{s \in \mathbb{S}}^{T} R_s^{(i^*)} \tag{2.11}$$

where $\mathbb{S}$ is the set of all consecutive pairs of long positions $\mathcal{L}_s \triangleq \{back, lay\}_s$ and/or short positions $\mathcal{S}_s \triangleq \{lay, back\}_s$. The trader could therefore have three different strategies; one of only long positions $\mathbb{S} = \{\mathcal{L}_1, \mathcal{L}_2, \ldots \mathcal{L}_{N_L}\}$ where N_L is the number of $\{back, lay\}$ pairs, one of only short positions $\mathbb{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots \mathcal{S}_{N_S}\}$ where N_S is the number of $\{lay, back\}$ pairs and a combination of short and long positions $\mathbb{S} = \{\mathcal{L}_1, \mathcal{L}_2, \ldots \mathcal{L}_{N_L}, \mathcal{S}_1, \mathcal{S}_2, \ldots \mathcal{S}_{N_S}\}$. The trader deploying this strategy would want Eq. (2.11) to be positive hence they would need to have more positive long and short positions than they have negative.

2.2 Horse Racing Data

The data sets used in this study came from the Betfair website [20], was sourced from [3] and went through many stages of cleaning and organising. The most substantial filtering that was done was the extraction of the in-play data. The

*Notice that we denote the winning horse here as i^* and the this index value is not known until time $t = T$.

in-play data is any data collected from the point when the race had started to the point of when it had finished. The important data fields are described in Table 2.1 but this section will give more details of the data fields.

The in-play data is split into individual races, and these are uniquely identified by an integer which is assigned to the race called the *EventID*; the race starting time and date is recorded as a string and is called *RaceStartDateTime*; different races have different format such as the type of horses competing, and this information is stored as a string referred to as *RaceType* (this is not used in this study); the individual horses in a race are identified either with the *HorseID* field which is an integer or the *HorseName* which is a string (both of which are unique to the selection or competing horse); each horse in the race has the time-dependent volume of total amount matched which has the units £: this is the total number of backs and lays matched where backs are positive and lays are negative (this is referred to as *TotalMatched*[*]); each horse in the race has a *Last Price Matched* signal referred to as *LPM* which are the final odds that were exchanged on a selection at a particular time.

The next fields are concerned with the limit-order-book: the back side of the book (the left side from the blue boxes in Fig. 2.2) are referred to as *BP1*, *BP2* and *BP3*, each of the three levels of the back prices have corresponding volumes referred to as *BV1*, *BV2* and *BV3* respectively all with the units £; the lay side of the book (the right side of the pink boxes in Fig. 2.2) are referred to as *LP1*, *LP2* and *LP3* and each of the three levels of the lay prices have corresponding volumes *LV1*, *LV2* and *LV3* respectively, all with the units £. The race time has been converted to an integer value in units of approximately ≈ 1.0 second and is referred to as *RaceTimeCS*; the mid-price is calculated as (BP1+LP1)/2 and called *MP* and the final data field is *Distance* which is the race distance from start to finish in miles. Some of the statistics of these data fields in Table 2.1 are discussed in Sec. 2.3. It is also worth stating that all fields were in decimal odds originally: LPM, BP1, BP2, BP3, LP1, LP2 and LP3 have been converted to implied probability by taking the reciprocal.

[*] This quantity rises when back odds are taken and falls if lay odds are taken.

Table 2.1: Data Key

Data Field	Data Type	Description
EventID	int	Unique identifier for individual race
RaceStartDateTime	str	Timestamp of the race starting time
RaceType	str	Format of the race
HorseID	int	Unique identifier for individual horses
HorseName	str	Unique identifier for individual horses
TotalMatched	double	Total value matched on horse (in £)
LPM	double	last price matched on horse
BP1	double	Back price 1 (back odds~buy price)
BP2	double	Back price 2 (back odds~buy price)
BP3	double	Back price 3 (back odds~buy price)
BV1	double	Back volume 1 (in £)
BV2	double	Back volume 2 (in £)
BV3	double	Back volume 3 (in £)
LP1	double	Lay price 1 (lay odds~sell price)
LP2	double	Lay price 2 (lay odds~sell price)
LP3	double	Lay price 3 (lay odds~sell price)
LV1	double	Lay volume 1 (in £)
LV2	double	Lay volume 2 (in £)
LV3	double	Lay volume 3 (in £)
RaceTimeCS	int	In-play racing time ≈ 1.0 sec
MP	double	Mid-price
Distance	double	Distance of the race in miles

2.2.1 Typical Race

Before delving into the statistics of the in-play horse racing signals let us gain insight into the dynamics of the variables in the system by visualising a typical set of signals. The set of race data that is presented is a race that took place on $2012 - 01 - 01$ and started at 14:20:00.0. Fig. 2.4 illustrates a standard set of in-play signals from the data such as LPM, Total Matched, order imbalance and spread. The middle left plot displays the price dynamics of the competing horses which come from the data field in Table 2.1 called LPM. The LPM values have been converted to implied probability from decimal odds. The price signals fluctuate more rapidly around the time $t \approx 0.7$. This increased activity is down to the fact that the market weighs information received closer to the end of the race greater than at the start. The market perceives information in this way because the certainty of the outcome is becoming more apparent, see Sec. 2.3.1. Notice that the price of the winning horse converges to a value just less than 1.0 or in decimal odds 1.01 and the price signals of the losing horses converge to a value close to 0.0 or in decimal odds 1000.0. The top left plot in Fig. 2.4 is the over-round, and this is defined as

$$OR_t \triangleq \sum_{i=1}^{N_h} LPM_t^{(i)} \tag{2.12}$$

where $LPM_t^{(i)}$ is the last price matched at time t and i is the index of the competing horses. The over-round is a measure of the fairness of the market in relation to selling or buying odds. Expanding on this point, if the over-round is greater than one, the market is a selling market (more lays than backs) as the book would be a Dutch one and vice versa for when the over-round is less than one. It is observed that this fluctuates around the mean value of 1; as a note, the over-round signal from other races has been observed having an upward drift with time. Also, the over-round has a volatility cluster that coincides with the LPM signal's volatility cluster at around $t \approx 0.7$.

The top right plot in Fig. 2.4 shows how the volume matched in-play changes in time. We see from this that the odds are not calculated with a parimutuel betting system as the red and purple LPM should then not cross each other. To see why this follows, consider the definition of the winning implied probability in the parimutuel betting system, Eq. (2.2), and compare this to the volumes matched in the top right plot in Fig. 2.4. The volume matched has an upward trend. If a back is matched the signal moves up, and down if a lay is matched, hence an upward trend indicates more backs are matched then lays. The order imbalance is a measure of the proportions of volumes on the back and lay side in the limit-order-book and is defined as

$$\rho_t^{(i)} \triangleq \frac{V_{t,B}^{(i)} - V_{t,L}^{(i)}}{V_{t,B}^{(i)} + V_{t,L}^{(i)}} \tag{2.13}$$

where $V_{t,B}^{(i)} = BV1_t^{(i)} + BV2_t^{(i)} + BV3_t^{(i)}$ and $V_{t,L}^{(i)} = LV1_t^{(i)} + LV2_t^{(i)} + LV3_t^{(i)}$: $\left\{BV1_t^{(i)}, BV2_t^{(i)}, BV3_t^{(i)}\right\}$ are the back volumes and $\left\{LV1_t^{(i)}, LV2_t^{(i)}, LV3_t^{(i)}\right\}$ are the lay volume variables both of which are described in the Table 2.1.

The bottom left frame in Fig. 2.4 shows that order imbalance, calculated as Eq. (2.13), is a variable that is dominated by noise throughout the race. This noise indicates that the order-book is fluctuating between a book that is heavy on the lay-side to one that is heavy on the back-side. When the market believes that the selection is not going to win the order imbalance converges to the value $\rho_t^{(i)} = 1$ and for the winning horse $\rho_t^{(i^*)} = -1$. The reason for this is: if it is believed a horse is never going to win the race then the market would only offer to sell backs as this will likely lead to positive cashflow, if this horse does not win. With regard to the winning selection the imbalance converges slowly to -1 if the race is competitive and fast if the race is not competitive.

The bottom right plot in Fig. 2.4 is the spread which is defined as the difference between the most competitive back and lay

$$\delta_t^{(i)} = b_t^{(i)} - l_t^{(i)} \tag{2.14}$$

where the best back is $b_t^{(i)} = \min\left\{BP1_t^{(i)}, BP2_t^{(i)}, BP3_t^{(i)}\right\}$ and the best ask is $a_t^{(i)} = \max\left\{LP1_t^{(i)}, LP2_t^{(i)}, LP3_t^{(i)}\right\}$. The sets $\left\{BP1_t^{(i)}, BP2_t^{(i)}, BP3_t^{(i)}\right\}$ and $\left\{LP1_t^{(i)}, LP2_t^{(i)}, LP3_t^{(i)}\right\}$ are the back and lay prices quoted in the order-book, described in the Table 2.1, which have been converted to implied probability. If one was working with decimal odds the minimum and maximum would be the other way round. The spread is strictly positive $\delta_t^{(i)} > 0$ which is observed in the bottom right plot Fig. 2.4.

The spread typically has a volatility cluster that coincides in time with the price and over-round volatility cluster. The spreads for the horses that are competing to win the race (red and purple) widen as each side of the book is exhausted and replenished. These fluctuations also indicate that the order-book prices are

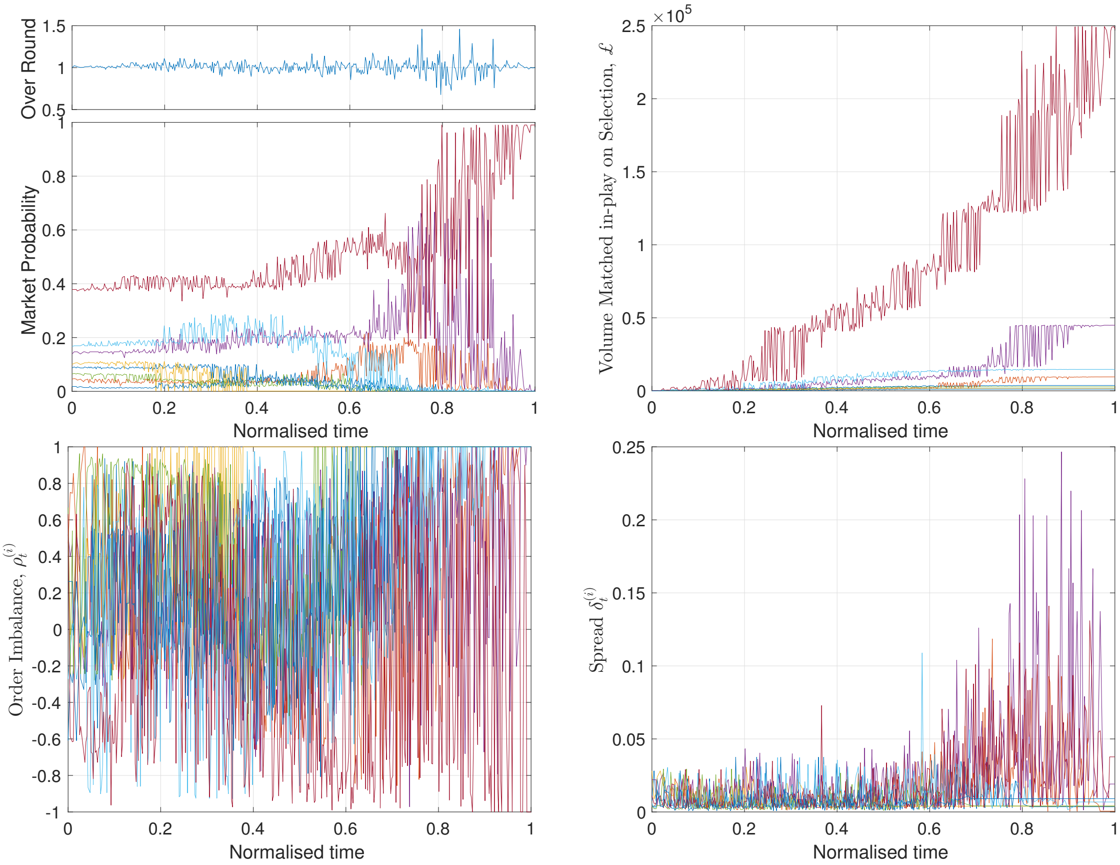


Figure 2.4: Examples of signals generated in a typical racing market, the race took place on 2012-01-01 and started at 14:20:00.0. Each different coloured line represents one of the eight competing horses in the race. Top left plot: this is the race over-round which is the sum of the implied probabilities of all the competing horses at each time step, see Eq. (2.12). Left middle plot: this is the last price matched of each horse competing in the race plotted against normalised time. Right top plot: this is the volume matched in £ on each selection where the total matched out-of-play has been removing centring the signal at the zero volume by subtracting the initial volume matched. Left bottom plot: these are the order imbalance signals against normalised time. Right bottom plot: this is the spread as calculated as in Eq. (2.14). One can observe that the order imbalance is a very noisy signal and the spreads tend to increase when the volatility in the price increases.

oscillating from a situation that is dominated by back orders to one dominated by sell orders.

2.3 Statistics of Data Set

The Table 2.2 gives a breakdown of the in-play racing data that is available to this study. There is a total of 12736 races which have been collected through Betfair [20]. The races occurred in the period starting on 31-12-11 and finishing on 31-12-12 and are located at race tracks across the British Isles. The total number of competing horses is 113999 with the smallest number in a particular race being 2 and the largest being 32, with an average of 8.95 horses. Note that the number 113999 is the number of LPM signals in the dataset and not the physical number of actual horses, as the same horse can compete in multiple races. The average total matched per time step $\Delta t \approx 1.0$ seconds in this same period is £738.28 seconds^{-1}, showing that the regular punter in these markets can get matched on small enough backs as on average the markets exhibit enough liquidity.

Table 2.2: Racing data

Data Field	Total	Min	Max	Mean	Standard Deviation
# of Races	12736	N/A	N/A	N/A	N/A
# of Horses	113999	2	32	8.95	3.23
TotalMatched, £	9.83×10^9	3.46×10^4	1.31×10^7	7.71×10^5	3.78×10^5
TotalMatched in-play, £	2.15×10^9	1.29×10^4	8.03×10^5	1.69×10^5	8.84×10^4
TotalMatched in-play, £ second^{-1}	9.40×10^6	83.22	$2,794.66$	738.28	240.49
Average Over-Round	N/A	0.69	1.55	1.02	0.049
Average Δt	N/A	1.03	4.47	1.15	0.09
# of Time Points	3006245	94	1099	236.043	111.425
Distance	N/A	0.625	4.125	1.5465	0.7909

The Table 2.3 displays the sample statistics of the increments of the last price matched signals for each of the selections. Each variable in this table is the result of filtering (or in the case of the All variable no filtering is applied). The in-play price increment signal is defined as

$$\Delta LPM_t^{(i)} = LPM_{t+1}^{(i)} - LPM_t^{(i)} \tag{2.15}$$

where i labels the competing horse and t are all the time steps within the races. The Winner, Favourite and Long-Shot* are created through the process of filtering (as they have fewer data points than the All variable); they represent the price increment signals of the horses that won the race, the horses with highest initial implied probability and the horses with the lowest initial implied probability respectively.

One can see from the Table 2.3 that the mean of the increments in absolute value is smallest in the All and Winner. The All variable has the most data points which is the reason it is an order of magnitude smaller than the long-shot and favourite mean. It is surprising that the mean of the Winner variable is the same order of magnitude as the All variable and in addition found also to be negative. This means on average the LPM signals move down but we know that at some point the LPM signal needs to move up toward the value ≈ 1.0. Hence, the increments that move the LPM signal up must on average occur less often than the down movements indicating the market must take a while to realise that the actual winner is going to win. The higher average increments found for the Favourite compared with the Long-Shot is down to the fact that the Favourite horses are generally competing for the pole position and actually win ≈ 34.66% of the time (see Fig. 2.4), resulting in a higher average number of up movements. The highest standard deviation is observed in the Favourite signals which makes sense as these signals are more likely to fluctuate when competing for the pole

Table 2.3: Summary of the statistics for the race increment signals of the LPM for the following selections: All, Winning, Favourite and Long-Shot. Each variable was found have a minimum value of -0.9890 and maximum value of 0.9891.

Variable	**Data Pts.**	**Mean**	**Std. Dev.**	**Skew.**	**Kurt.**
All	26067165	5.74×10^{-6}	0.098	0.096	32.6
Winner	2993509	-5.15×10^{-6}	0.102	0.072	29.71
Favourite	2993509	5.58×10^{-5}	0.148	0.019	11.9
Long-Shot	2993509	1.86×10^{-5}	0.053	0.251	141.3

*The long-shot is the horse or horses with the smallest implied probability of winning the race.

position and the Long-Shot signals less likely. The skewness for all the variables

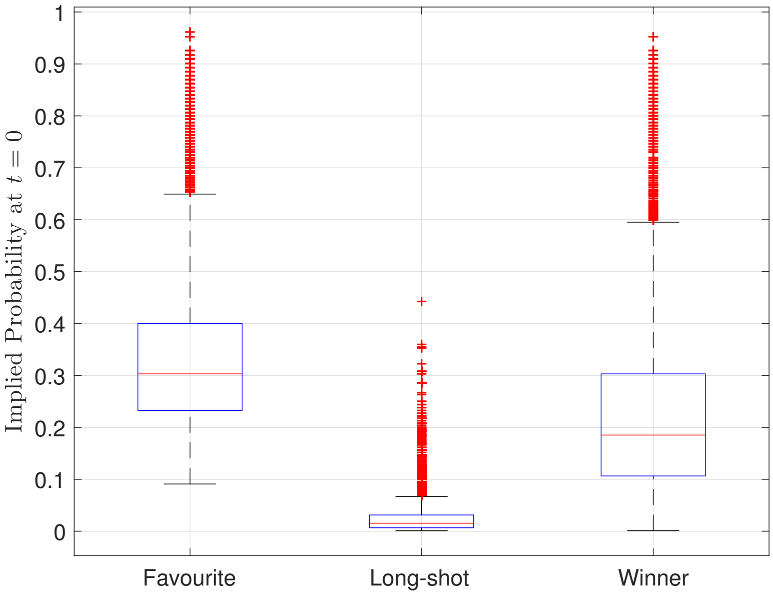


Figure 2.5: These box-plots represent the initial odds quoted at the first time point. The Favourite is the selection with the highest implied probability, the Long-Shot is the selection with the lowest implied probability and the Winner is the selection that actual won the race. For each box-plot there are 2993509 data points. The median of the initial odds is represented by the red line and the top and bottom of the blue box are the 75th and 25th percentiles respectively. The whiskers extend to the upper and lower extreme quantiles as defined by the Tukey range test [1] and the red crosses are the outlying data points.

is found to be positive but the largest is the Long-Shot: indicating that all the variables have a distribution which is longer in the right tail and has more mass concentrated in the left. The kurtosis values for all variables are greater than the normal distribution, indicating a positive excess kurtosis and hence a leptokurtic distribution. The highest kurtosis is found for the Long-Shot resulting in a distribution with the strongest leptokurtic features and this could also be the reason for the smallest standard deviation.

Fig. 2.5 shows the box-plots of all the initial LPM odds posted at $t = 0$ which are the odds quoted when the race has just started. The odds have been converted from decimal odds to implied probability. This figure displays the initial odds for the selections that correspond to the Favourite, Long-Shot and Winner, where these definitions are the same as previously stated in Table 2.3. One can observe the intuitive result that the mean of the Favourite is greater than the Winner and Long-Shot, since a favourite is always given the highest initial implied probability of winning. The interquartile range, which is the distance between the 75th and 25th quantiles, is found to be larger for the Favourite than for the Long-Shot. This smaller interquartile range for the Long-Shot as compared to the Favourite shows that there are different risk aversion preferences for the market agents who trade the Long-Shots compared the Favourites. This would be an interesting statistical property to explore from the point-of-view of behavioural finance and decision making under uncertainty, one such model being prospect theory [39] but this question is left open.

Table 2.4: True unconditional probabilities (This means the actual probability of the Favourite/Long-Shot winning the race, see Eq. (2.16)) and empirical market LPM probabilities denoted as $P_{t=0}$ for Favourite and Long-Shot horses. The asterisk $\ast$ means the statistics where generated via a resampling bootstrap with the number of resamples 1.0×10^5.

Variable	Estimates	Std. Dev.
$\mathbb{P}\left(\text{Favourite} = \text{Winner}\right)$	0.3466	0.0042^{*}
$\mathbb{P}\left(\text{Long-Shot} = \text{Winner}\right)$	0.0268	0.0014^{*}
$\langle \max\left\{P_{t=0}\right\}\rangle_{N_r}$	0.3321	0.1387
$\langle \min\left\{P_{t=0}\right\}\rangle_{N_r}$	0.0256	0.0325

The statistical spread for the initial implied probability (the Std. Dev. field in Table 2.4) for the Winner is also larger than the Favourite which comes down to the fact that the Favourite does not always prove to be the winner, in fact for this set of data the Favourite has an unconditional chance of winning 0.3466, see Table 2.4 and calculated with Eq. (2.16). The eventual winner will be a mix of long-shot, non-favourites and favourites which effectively increases the standard deviation of the odds. The long-shot odds has the smallest standard deviation meaning the market prices such selections closer to the unconditional probability (see Table 2.4) where the long-shot wins with a 0.0268 probability. In each of the box plots in Fig. 2.5 there are outliers, which are defined from Tukey's test [1]. Some of these points are outliers because the data points are a collection of all races that contain different numbers of horses, for example, a race with two horses will weight the long-shot higher than a race with 32 horses. The Table 2.4 highlights some differences that emerge from the market's estimation of implied probability compared to the true unconditional probabilities. The true unconditional probability of the Favourite winning is defined as

$$\mathbb{P}\left(\text{Favourite} = \text{Winner}\right) = \frac{1}{N_r}\sum_{j=1}^{N_r} \mathbb{1}_{\left\{\text{Favourite}_{t=0}^{(j)}=\text{Winner}_{t=T}^{(j)}\right\}} \tag{2.16}$$

where $\mathbb{1}$ is the indicator which is 1 when the favourite horse from race j is the eventual winner and 0 otherwise. Using the race data, we estimated Eq. (2.16) to have the value $\mathbb{P}\left(\text{Favourite} = \text{Winner}\right) = 0.3466 \pm 0.0021$ which is close to the market's average estimation of the initial odds $\left\langle \max\left\{P_{t=0}^{(i)}\right\}\right\rangle_{N_r} = 0.3321 \pm 0.0694$ where N_r is the number of races averaged over, note that $P_{t=0}^{(i)}$ denotes the initial implied probability of horse i. This closeness of true unconditional probabilities and the average implied probabilities means that the market on average is almost correctly estimating the odds of a Favourite or Long-Shot winning. We will return to this in Sec. 2.4 where a model for the order statistics of the initial odds is developed to test how well the market ranks the horses.

2.3.1 Statistical Dispersion of the In-play Odds

Statistical dispersion is a class of measures used to evaluate the spread* or relative spread of data or to put it another way, how stretched or squeezed an empirical distribution is. The most common of these is the sample standard deviation. The implied probabilities as defined in Eq. (2.1) are decided by the market and can be considered a representation of a price. This section will explore different measures of the variability in the implied odds as a function of time. The measures used are the following: Gini index, Theil index, Atkinson index, and Generalised entropy index. This set of measures show how uncertainty/certainty evolves in horse racing markets.

*The term spread in this subsection should not to be confused with the spread defined in Eq. (2.14) – we mean statistical spread – for example, standard deviation.

Each horse at a given time will have a market determined implied probability $P_t^{(i)}$ where $i = 1, 2, \dots, N_h$. The first statistical dispersion measure used is the *Gini Coefficient* [40], which is equivalent to half the value of the statistical measure known as the *relative mean absolute difference*. The *Gini Coefficient* for the set of prices $\left\{P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right\}$ is defined

$$G_t = \frac{\sum_{i=1}^{N_h} \sum_{j=1}^{N_h} \left|P_t^{(i)} - P_t^{(j)}\right|}{2N_h \sum_{j=1}^{N_h} P_t^{(j)}} \tag{2.17}$$

where the time $t \in [0, 1]$ is the normalised time and N_h is the number of horses in each race. The Gini coefficient is not a coherent measure – because the absolute value makes this non analytic – which is defined to satisfy the four following conditions: monotonicity, sub-additivity, homogeneity, and translational invariance [41]. As a result it is not additive, which leads to the issue that one cannot simply find the average of two Gini coefficients as it would be inconsistent with the actual Gini coefficient of a two part system. Alternative measures used that are coherent are the entropy measures also known as the redundancy measures.

The three entropic measures that will be discussed have no apparent merit over one another, but mathematically they are all additive and coherent measures, which is a clear advantage over the Gini index. The Theil index [42], also known as the Von Neumann entropy [43], is the first redundancy measure and is defined as

$$T\left(P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right) = \frac{1}{N_h} \sum_{i=1}^{N_h} \frac{P_t^{(i)}}{\mu_t} \ln\left(\frac{P_t^{(i)}}{\mu_t}\right) \tag{2.18}$$

where μ_t is the time dependent sample mean of the prices $\mu_t = \frac{1}{N_h} \sum_{i=1}^{N_h} P_t^{(i)}$. The second measure is the Atkinson Index [44] which is defined as

$$A_\varepsilon\left(P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right) = \begin{cases} 1 - \frac{1}{\mu_t}\left(\frac{1}{N_h} \sum_{i=1}^{N_h} \left(P_t^{(i)}\right)^{1-\varepsilon}\right)^{1/(1-\varepsilon)} & \text{for } 0 \leq \varepsilon \neq 1 \\ 1 - \frac{1}{\mu_t}\left(\prod_{i=1}^{N_h} P_t^{(i)}\right)^{1/N_h} & \text{for } \varepsilon = 1, \end{cases} \tag{2.19}$$

where the $\varepsilon \in (0, \infty)$ is the measures' parameter that models the level of affinity a set of prices exhibit toward being diverse (spread out) or non-diverse (close together). For example if $\varepsilon = 0$ there is an affinity for the prices to be non-diverse and if $\varepsilon = \infty$ there is an affinity for price to be diverse. The Generalised entropy index [45] is defined as

$$GE_{\alpha}\left(P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right) = \begin{cases} \frac{1}{N_h \alpha(\alpha-1)} \sum_{i=1}^{N_h} \left[\left(\frac{P_t^{(i)}}{\mu_t}\right)^{\alpha} - 1\right], & \alpha \neq 0, 1, \\ \frac{1}{N_h} \sum_{i=1}^{N_h} \left(\frac{P_t^{(i)}}{\mu_t}\right) \ln\left(\frac{P_t^{(i)}}{\mu_t}\right), & \alpha = 1, \\ -\frac{1}{N_h} \sum_{i=1}^{N_h} \ln\left(\frac{P_t^{(i)}}{\mu_t}\right), & \alpha = 0. \end{cases} \tag{2.20}$$

Notice that the Theil index Eq. (2.18) and Atkinson index Eq. (2.19) are defined within the mathematical generalisation of Eq. (2.20), such that

$$\begin{aligned} T\left(P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right) &= GE_{\alpha=1}\left(P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right) \\ A_{\varepsilon}\left(P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right) &= 1 - \exp\left(GE_{\alpha}\left(P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right)\right) \end{aligned} \tag{2.21}$$

where $\varepsilon = 1 - \alpha$. Using the four measures Eq. (2.17) - (2.20) on the set of prices $\left\{P_t^{(1)}, P_t^{(2)}, \dots, P_t^{(N_h)}\right\}_{j=1,2,\dots N_r}$ where $N_r = 12736$ (the number of races in data set) gives the statistical dispersion estimations for each for the uncertainty measures. The results are binned into 40 equally spaced bins in the time period $t \in (0, 1)$ and averaged in those bins for each measure. The results are displayed in Fig. 2.6. One can see that the results that are averaged over the number of races N_r monotonically increase. The structure of the curves in Fig. 2.6 shows that the uncertainty in price on average is always decreasing as time evolves. This result backs up our intuition as we know that the prices in typical races (see Fig. 2.4) have a common terminal structure. That is when a race starts the prices have maximum diversity as the market is uncertain about which horse will win and as the race draws to its conclusion the uncertainty decreases and so does the spread of prices.

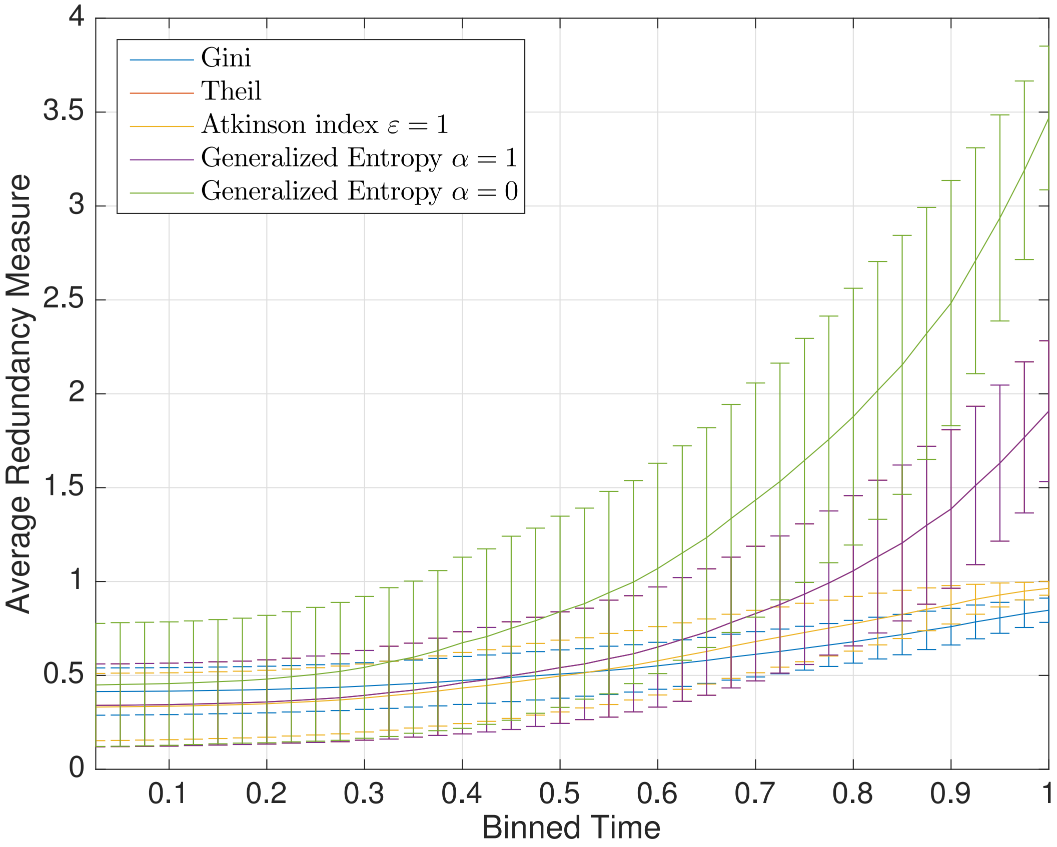


Figure 2.6: Five different sample averages of the parametrised statistical dispersion measures as indicated in the legend. All are monotonically increasing and the Theil measure is equivalent to the generalised entropy $\alpha = 1$. As these measures are monotonic, the dispersion of the average in-play odds is always increasing.

2.4 Order statistics of horse racing gambling and random divisions on an interval

This section looks to build a model for the distribution of initial odds as quoted by the in-play gambling markets, see Fig. 2.5, and the results of which where presented in [4].

The idea is to assume that the implied probabilities quoted by the market sum to one, in this case, the last price matched are used.

This assumption is not too far away from the truth, but we know from Sec. 2.2 that the sum in the in-play markets can diverge above or below unity. This effect is not as prominent in the initials odds.

The segments on a unit interval $[0,1]$ are constructed to represent the implied probabilities – if it is assumed that there is little to no information on the horses' form and past performance. The simplest model for the distribution of these segments that one could propose is drawing $N_h - 1$ samples from a uniform distribution on the unit interval. This model can be interpreted as the randomly broken stick problem [46], where the stick is of a unit length. The largest segment created by this sampling is construed as the odds of the favourite horse, the second biggest as the second favourite, and so on until the smallest segment, which is the odds of the long-shot.

Using the well known result of the mapping between exponential random variables and the statistics of the random division, one can hypothesise that the horses' and jockeys' joint *abilities* are exponentially distributed. If such abilities are distributed in this way then prices on average, in economic sense, are efficient as the no market agent can find a more accurate price than the market price and all agents in market will price according to this knowledge. Hence, the probability of the j^{th} horse winning the race is directly proportional to this ability and it can be assumed that

$$\mathbb{P}\left[j^{\text{th}} \text{ horse wins}\right] \triangleq \mathcal{P}_j = \frac{A_j}{\sum_{i=1}^{N_h} A_i} \tag{2.22}$$

where $A_i \sim \text{Exp}(\lambda)$ where $\lambda^{-1} = \frac{1}{N_h}\sum_{i=1}^{N_h} A_i$ and $i = 1, 2, \ldots, N_h$. The j^{th} segment is denoted as U_j and defined to be $U_j \triangleq \mathcal{P}_j$, therefore bounded by the randomly partitioned interval $[0,1]$. Ordering the segments U_j by their lengths is a procedure denoted as $U_{(k)}$ where (k) is the k^{th} largest segment, see fig 2.7. Thus, the hypothesis that the horses' ability is exponentially distributed, defined in Eq. (2.22), is represented as the following expression

$$\mathcal{P}_{(k)} = U_{(k)} \tag{2.23}$$

where $k = 1, 2, \ldots, N_h$ and the ordering of the segments is shown in Fig. 2.7. The statistical testing of this hypothesis is the central question in this section. To put this in to the context of horse gambling markets: when a bookie says they have "*a hot tip*" and their odds are better than the market odds this is not true if and only if Eq. (2.23) holds true – to some level of statistical confidence. Such a test gives *precisely* a measure of the degree of price efficiency in initial quoted odds in the horse gambling markets. To give more detail of what one means by testing the efficiency: consider a gambler with sufficient information such that they can *correctly* rank/order the selections, but this information is very noisy so they can not correctly assign the exact probability of the horses winning the race. Therefore, we test how efficiently the market is at ranking the horses on average.

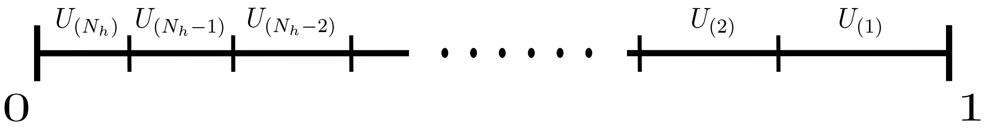


Figure 2.7: The unit interval partitioned N_h times into segments that are ordered such that $0 < U_{(N_h)} < U_{(N_h-1)} < U_{(N_h-2)} < \cdots < U_{(2)} < U_{(1)}<1$.

It is known from [21] (page 135) that the segments created by partitioning the unit interval by a exponential distribution, as shown in Fig. 2.7, has the following expectation for the k^{th} largest segment

$$\bar{U}_{(k)} = \frac{1}{N_h}\left(\frac{1}{N_h} + \frac{1}{N_h - 1} + \cdots + \frac{1}{k}\right) = \frac{1}{N_h}\sum_{j=k}^{N_h}\frac{1}{j} \tag{2.24}$$

where the bar represents the statistical average for a given N_h, for example $\bar{X} = \mathbb{E}\left[X \mid N_h\right]$. In this case one is expecting that on average the initial odds of the k^{th} highest ranked horse is equivalent to Eq. (2.24). It is trivially found that the expectation of the smallest segment is

$$\bar{U}_{(N_h)} = \frac{1}{N_h^2} \, . \tag{2.25}$$

The harmonic function can be used to represent the sum on the right hand side of Eq. (2.24) and is defined as

$$H_{N_h,k} = \sum_{j=k}^{N_h} \frac{1}{j} \, , \tag{2.26}$$

where one will notice that $\bar{U}_{(N_h)} = \frac{1}{N_h} H_{N_h,N_h}$ is the expectation of the smallest segment, Eq. (2.25), and $\bar{U}_{(k)} = \frac{1}{N_h} H_{N_h,k}$ is the expectation of the k^{th} largest segment, Eq. (2.24). If the initial implied probability quoted by the market are exponentially distributed, where $P_{(k)}$ denotes k^{th} favourite horse, then it follows from the hypothesis Eq. (2.23) that

$$P_{(k)} = \bar{U}_{(k)} \, . \tag{2.27}$$

Table 2.5 presents the estimates for the average initial odds quoted in the market $P_{(k)}$ and the true winning probabilities of the horse winning (calculated as in Eq. (2.16)) and the expected largest segment Eq. (2.24). These three statistics are compared for the favourite $k = 1$, second favourite $k = 2$, third favourite $k = 3$, fourth favourite $k = 4$, and the long-shot $k = N_h$. In Table 2.5 we compare the averages for each subsets: all N_h, $N_h \leq 7$, $8 \leq N_h \leq 10$ and $N_h \geq 11$. Finding that the estimates of these conditional averages, especially for the top ranked horses, are in agreement with empirical market odds. We therefore conclude, to the first order, that there is statistical truth in the relationship Eq. (2.23) holding; and thus there is some level of price efficiency in the market value of initial odds.

	favourite	2^{nd} favourite	3^{rd} favourite	4^{th} favourite	long-shot
	$k = 1$	$k = 2$	$k = 3$	$k = n$	
$\langle P_{(k)} \rangle$	0.3321	0.2048	0.1429	0.1025	0.0256
$\langle \mathcal{P}_{(k)} \rangle$	0.3478	0.2047	0.1348	0.0952	0.0281
$\langle \bar{U}_{(k)} \rangle$	0.3380	0.2089	0.1447	0.1033	0.0197
	favourite	2^{nd} favourite	3^{rd} favourite	4^{th} favourite	longshot
	$k = 1$	$k = 2$	$k = 3$	$k = n$	
$\langle P_{(k)} \mid N_h \leq 7 \rangle$	0.4173	0.2460	0.1576	0.0990	0.0442
$\langle \mathcal{P}_{(k)} \mid N_h \leq 7 \rangle$	0.4358	0.2423	0.1484	0.0853	0.0402
$\langle \bar{U}_{(k)} \mid N_h \leq 7 \rangle$	0.4332	0.2462	0.1538	0.0958	0.0377
	favourite	2^{nd} favourite	3^{rd} favourite	4^{th} favourite	longshot
	$k = 1$	$k = 2$	$k = 3$	$k = n$	
$\langle P_{(k)} \mid 8 \leq N_h \leq 10 \rangle$	0.3184	0.1985	0.1438	0.1078	0.0182
$\langle \mathcal{P}_{(k)} \mid 8 \leq N_h \leq 10 \rangle$	0.3327	0.2081	0.1361	0.1031	0.0233
$\langle \bar{U}_{(k)} \mid 8 \leq N_h \leq 10 \rangle$	0.3166	0.2041	0.1478	0.1103	0.0128
	favourite	2^{nd} favourite	3^{rd} favourite	4^{th} favourite	longshot
	$k = 1$	$k = 2$	$k = 3$	$k = n$	
$\langle P_{(k)} \mid N_h \geq 11 \rangle$	0.2470	0.1631	0.1247	0.1004	0.0119
$\langle \mathcal{P}_{(k)} \mid N_h \geq 11 \rangle$	0.2614	0.1564	0.1172	0.0977	0.0193
$\langle \bar{U}_{(k)} \mid N_h \geq 11 \rangle$	0.2500	0.1703	0.1305	0.1039	0.0065

Table 2.5: Using the in-play horse racing market data discussed in Sec. 2.2: we filter the initial odds (last price matched) and sort the implied odds in rank order favourite, 2^{nd} favourite, 3^{rd} favourite, $\ldots$, long-shot. Denoting the following k^{th} largest variables: the implied market odds $P_{(k)}$, the winning probabilities $\mathcal{P}_{(k)}$ estimated simpler to Eq. (2.16), and the expected segment lengths $\bar{U}_{(k)}$ calculated as Eq. (2.24). The sample average is denoted $\langle . \rangle$ and estimated for the average for all N_r races and sub-samples with the number of horses $N_h \leq 7$, $8 \leq N_h \leq 10$, $N_h \geq 11$. The theoretical expectation of the segment lengths are calculated by averaging the first moment of Eq. (2.30) over the empirical distribution of N_h.

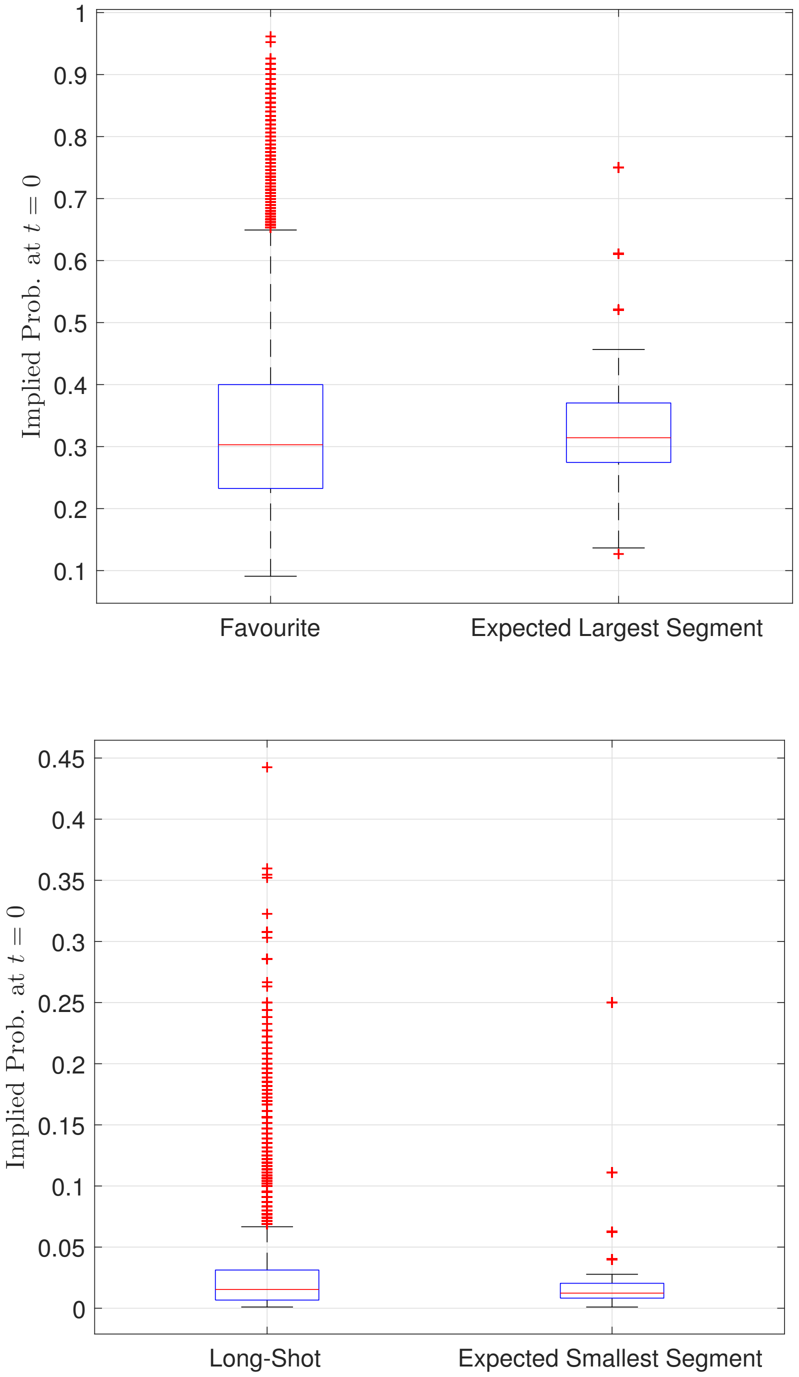


Figure 2.8: We compare the distributions of the implied market odds $\mathcal{P}_{(k)}$ and the expected segment lengths $\bar{U}_{(k)}$, calculated as Eq. (2.24), where $k = 1$ (favourites) and $k = N_h$ (long-shots). Top Plot: A box plot comparison between the initial favourite implied probabilities and the expected largest segment $\frac{1}{N_h} H_{N_h,1}$. Bottom Plot: A box plot comparison between the initial long-shot implied probabilities and the expected smallest segment $\frac{1}{N_h^2}$.

From the Tables 2.4 and 2.5 one observes that the average found from $N_r = 12736$ races with the unbiased measure Eq. (2.24) for the expectation of the largest segment is $\left\langle \bar{U}_{(1)} \right\rangle_{N_r} = 0.338 \pm 0.045$. This result is not far from the unconditional probability $\left\langle \mathcal{P}_{(1)} \right\rangle_{N_r} = 0.346 \pm 0.002$ and the empirical value $\left\langle P_{(1)} \right\rangle_{N_r} = 0.332 \pm 0.069$, with the absolute difference between these empirical values and model calculated as 0.009 and 0.004 respectively. One can observe from the top panel in Fig. 2.8 that the interquartile range of the box plot for the largest expected segment $\frac{1}{N_h} H_{N_h,1}$ is within the interquartile range of the empirical favourite odds $P_{(1)}$. The smallest expected segment is $\left\langle \bar{U}_{(N_h)} \right\rangle_{N_r} = 0.02 \pm 0.02$ and comparing this with $\left\langle \mathcal{P}_{(N_h)} \right\rangle_{N_r} = 0.027 \pm 0.001$ and $\left\langle P_{(N_h)} \right\rangle_{N_r} = 0.02 \pm 0.03$ we see again they are close. The bottom panel in Fig. 2.8 shows that the interquartile range for the box plot of the smallest expected segment $\frac{1}{N_h} H_{N_h,N_h}$ is within the interquartile range of the empirical long-shot odds $P_{(N_h)}$. The model $\frac{1}{N_h} H_{N_h,N_h}$ does not perform as well when the number of horses considered in the average is large.

For example $8 \leq N_h \leq 10$ and $N_h \leq 11$. We believe the model underperforms in this way because when the number of horses becomes large enough the market finds it difficult to rank the least favourite horses.

Defining the difference between the empirical value of the k^{th} favourites' initial odds and the unbiased estimator Eq. (2.24) as

$$D_{(k)} = P_{(k)} - \bar{U}_{(k)} = P_{(k)} - \frac{1}{N_h} H_{N_h,k}. \tag{2.28}$$

This measure $D_{(k)}$ gives a means of quantifying how much the empirical initial implied probabilities differ to the model Eq. (2.24), which can be tested to see if this difference has any correlation to the final result of the races. Defining the indicator $\mathbb{1}_{(k)} \in \{0, 1\}$, which is 1 when the k^{th} favourite wins and 0 otherwise. It can be shown that the estimate for the favourite is $\text{Corr}\left[D_{(1)}, \mathbb{1}_{(1)}\right] = 0.2296$ and long-shot is $\text{Corr}\left[D_{(N_h)}, \mathbb{1}_{(N_h)}\right] = 0.1251$, hence there is a small positive correlation which indicates that there exists a linear relationship between the divergence and the race outcome.

2.4.1 Order Statistics of the Random Division of the Unit Interval

Here we present the mathematics of the order statistics of the random division of the unit interval, which can be found in [21, 46]. We do this here so can find the theoretical distribution of the segments $U_{(k)}$ so we can plot and compare this to the empirical distributions.

Here we use a similar line of reasoning as in the book by Holst and Nagaraja [46], which can be found on page 135. The probability that any particular j segments have lengths simultaneously longer than $c_1, c_2, \ldots, c_j$, respectively (where $\sum_{i=1}^{j} c_i \leq 1$) is

$$\mathbb{P}[U_1 > c_1, U_2 > c_2, \cdots, U_j > c_j \mid N_h] = (1 - c_1 - c_2 - \cdots - c_j)^{N_h - 1} \tag{2.29}$$

which is proved in [21] Chapter 6. In order to have pieces with a length that is at least $c_1, c_2, \ldots, c_k$ for all $N_h - 1$, the cuts have to occur as segments within the unit interval $[0,1]$ with a total length $1 - c_1 - c_2 - \cdots - c_k$. Consider the example where $\mathbb{P}[U_1 > c_1 \mid N_h]$ is the probability that all $N_h - 1$ cuts occur in the interval $(c_1, 1]$, and since the cuts are randomly distributed in $[0,1]$ then $\mathbb{P}[U_1 > c_1 \mid N_h] = (1 - c_1)^{N_h - 1}$. Note that from [46] the complementary cumulative distribution function (CCDF) $\mathbb{P}[U_{(k)} > x \mid N_h]$ is found to be

$$\begin{aligned}\mathbb{P}[U_{(k)} > x \mid N_h] = &\sum_{j=1}^{k-1} \binom{N_h}{j} \sum_{\ell=0}^{N_h - j} (-1)^{\ell-1} \binom{N_h - j}{\ell} [1 - (j + \ell)x]_+^{N_h - 1} \\ &+ \sum_{\ell=1}^{N_h} (-1)^{\ell-1} \binom{N_h}{\ell} [1 - \ell x]_+^{N_h - 1} ,\end{aligned} \tag{2.30}$$

where $a_+ = \max[a, 0]$ and using the convention that $\sum_{j=1}^{0} = 0$. The length of the largest segment is thus distributed according to

$$\mathbb{P}[U_{(1)} > x \mid N_h] = \sum_{\ell=1}^{N_h} (-1)^{\ell-1} \binom{N_h}{\ell} [1 - \ell x]_+^{N_h - 1} . \tag{2.31}$$

The average length of the k^{th} largest segment is given by:

$$\begin{aligned}\bar{U}_{(k)} &= \int_0^1 \mathrm{d}x\ \mathbb{P}[U_{(k)} > x \mid N_h] \\ &= \frac{1}{N_h}\sum_{j=1}^{k-1}\binom{N_h}{j}\sum_{\ell=0}^{N_h-j}(-1)^{\ell-1}\binom{N_h-j}{\ell}\frac{1}{j+\ell} + \frac{1}{N_h}\sum_{\ell=1}^{N_h}(-1)^{\ell-1}\binom{N_h}{\ell}\frac{1}{\ell}\end{aligned} \tag{2.32}$$

leading to the result that

$$\bar{U}_{(k)} = \frac{1}{N_h}\sum_{j=k}^{N_h}\frac{1}{j} = \frac{1}{N_h}H_{N_h,k}\ , \tag{2.33}$$

which is Eq. (2.24). The quadratic variation found in [4] is given by

$$\begin{aligned}\overline{U^2}_{(k)} &= 2\int_0^1 \mathrm{d}x\ x\mathbb{P}[U_{(k)} > x \mid N_h] \\ &= \frac{2}{N_h(N_h+1)}\left[\sum_{j=1}^{k-1}\binom{n}{j}\sum_{\ell=0}^{N_h-j}(-1)^{\ell-1}\binom{N_h-j}{\ell}\frac{1}{(j+\ell)^2} + \sum_{\ell=1}^{N_h}(-1)^{\ell-1}\binom{N_h}{\ell}\frac{1}{\ell^2}\right]\ .\end{aligned} \tag{2.34}$$

which simplifies to

$$\overline{U^2}_{(k)} = \frac{2}{N_h(N_h+1)}\sum_{j=k}^{N_h}\frac{H_{N_h,j}}{j} = \frac{2}{N_h+1}\sum_{j=k}^{N_h}\frac{\bar{U}_{(j)}}{j}\ . \tag{2.35}$$

This result is used in Sec. 2.4.2 to calculate the probability that the k^{th} favourite horse wins is $\bar{U}_{(k)}$. Fig. 2.9 shows the empirical complementary cumulative density function (ECCDF) of the k^{th} favourite's implied odds where $k = 1, 2, 3, 4$ and N_h compared against the theoretical CCDF of the k^{th} longest segment, Eq. (2.30) with $k = 1, 2, 3, 4$ and N_h. The agreement between the ECCDF and CCDF is apparent but more statistical work is needed before this is confirmed with statistical confidence.

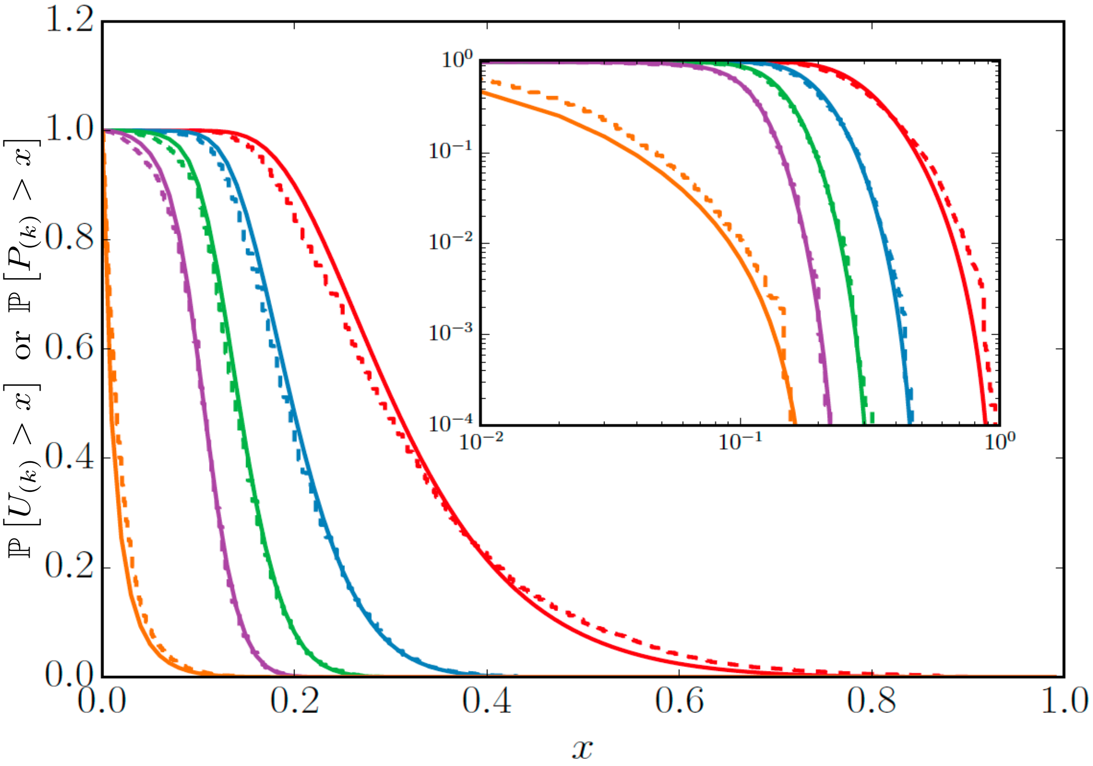


Figure 2.9: The dashed lines are the the empirical complementary cumulative density function (ECCDF) for the implied odds of the k^{th} favourite horses' denoted as $\mathbb{P}[P_{(k)} > x]$ and the solid lines are the theoretical complementary cumulative density function (CCDF) of the k^{th} largest segment of the division of the unit interval, which is Eq. (2.30). Note that $k = 1, 2, 3, 4,$ and N_h which are respectively the red, blue, green, purple and orange lines. We see for the favourite distributions (red line) that there are diverges in the empirical distribution (dashed red line) from the theoretical model (solid red line). Hence, the market does under price the favourite, a phenomenon known as the favourite-longshot bias [2], and over prices the races with a low number of race – which is seen for the probabilities where $x > 0.4$.

2.4.2 Odds of the Winning Horse

From Sec. 2.4.1 we showed that the probability that the k^{th} favourite horse wins is $\bar{U}_{(k)}$. The implied odds of the k^{th} horse given that it wins is a different estimation and from the viewpoint of the random division of the unit interval one can work this out. Consider that I choose a random point on the interval $[0,1]$ and given that the random point lies in the k^{th} largest segment, what is the expected length of this segment? Define the indicator $\mathbb{1}_{(k)} = 1$ if the random point lies in the k^{th} largest segment and $\mathbb{1}_{(k)} = 0$ else. Then the probability that the k^{th} horse given that it wins is

$$\mathbb{P}[U_{(k)} = x \mid \mathbb{1}_{(k)} = 1] = \mathbb{P}[\mathbb{1}_{(k)} = 1 \mid U_{(k)} = x]\frac{\mathbb{P}[U_{(k)} = x]}{\mathbb{P}[\mathbb{1}_{(k)} = 1]} = \frac{x\mathbb{P}[U_{(k)} = x]}{\bar{U}_{(k)}} \tag{2.36}$$

and the expectation is

$$\mathbb{E}[U_{(k)} \mid \mathbb{1}_{(k)} = 1] = \frac{\overline{U^2}_{(k)}}{\bar{U}_{(k)}} \,. \tag{2.37}$$

The average implied odds of the winning horse is estimated as $\bar{P}_{\text{win}} = 0.2148$. The average length of the segment containing the random point, which is assumes that the horses' abilities are exponentially distributed and the gambling market is efficient, thus

$$\left\langle \mathbb{E}[U_{(k)} \mid \mathbb{1}_{(k)} = 1] \right\rangle = \sum_{k=1}^{N_h} \mathbb{E}[U_{(k)} \mid \mathbb{1}_{(k)} = 1]\mathbb{P}[\mathbb{1}_{(k)} = 1] = \sum_{k=1}^{N_h} \overline{U^2}_{(k)} = \frac{2}{N_h + 1} \,. \tag{2.38}$$

which after averaging over N_r yields 0.2107, again very close to the empirical value. Table 2.6 compares the implied odds of the k-th favourite given that it wins with the average length of the k-th largest segment given that it contains a random point, see Eq. (2.37).

2.5 Conclusion and Summary

This chapter has given a detailed discussion of the online gambling industry and the markets that have been created because of its growth due to the internet boom. We discuss the main points highlighted by this study and the implications

	favourite $k = 1$	2^{nd} favourite $k = 2$	3^{rd} favourite $k = 3$	4^{th} favourite $k = n$	long-shot
$\left\langle P_{(k)} \middle\vert \text{win} \right\rangle$	0.3735	0.2148	0.1542	0.1139	0.0886
$\left\langle \bar{U}_{(k)} \middle\vert \mathbb{1}_{(k)} = 1 \right\rangle$	0.3622	0.2196	0.1549	0.1145	0.0383

Table 2.6: Average implied odds given that the horse wins and expected segment lengths given that it contains a random point for all races in our dataset (with $N_h \geq 5$). The theoretical expectation of the segment lengths are calculated by averaging Eq. (2.37) over the empirical distribution of N_h.

of the overall results.

A discussion is found in Sec. 2.1 where a breakdown of the market structure and the revenues highlights that online gambling is the fastest growing sector of the industry. The contrast of between gambling and investment is discussed, and some of the similarities are highlighted, which can blur the differences between these two concepts. The differences come down to the fundamental mechanism they play in society, but some of the strategies are very similar. The dataset used in this study came from a horse racing exchange, so details of this particular market are explained such as race selection, horse selection order-book, backing and laying odds. The exchanged assets in gambling markets are the odds which are auctioned (in a manner similar to but different from financial market order-books), and we discussed how to interpret such instruments. The hedging of such instruments is illustrated as a two positioning process where one backs and then lays, or vice versa and gains (or losses) are extracted from the change in the odds between the two positions.

In Sec. 2.2 the actual data set and various fields that constitute it are given a detailed description, discussing what they refer to in the market and also in the race itself, such the selection or event identification. It is also stated that this study concentrates its effort in the exploration of statistical properties that emerge from in-play markets. In Sec. 2.3 we presented Table 2.2 outlining the sample statistics of the data set such as the average market liquidity. The Table 2.3 described the moments of the price increment signals extracted from the dataset showing the distributions are not normal and a have highly leptokurtic structure. We also found some interesting statistics presented in Table 2.4 regarding the initial odds posted, where one observes that agents on average rank the selection close to the unbiased estimator which is shown in Table 2.5. If the market diverges away from the unbiased estimator, we find that there is a small positive correlation with outcome of the race, indicating that agents could be diverging from the unbiased estimator because of some extra information they have. Using statistical dispersion measures and entropic measures we observe the intuitive statistical fact that on average the dispersion in the odds increases, see Fig. 2.6.

In Sec. 2.4 we have found a remarkable agreement between the order statistics of the randomly broken stick and the statistical properties of horse racing betting markets. Because we observe the empirical values of the implied odds and true winning probabilities to be close, we conclude that this betting market is informationally efficient[*], at least to some degree. Discrepancies are found for the long-shot, suggesting that gamblers fail to rank the horses accurately when their number is significant. Assuming that the implied odds reflect to a large extent the true winning probabilities, we conclude that the "ability"[†] of a horse can be defined in such a way that its winning probability is the ratio of its "ability" to the sum of all its competitors' abilities, provided "ability" is exponentially distributed.

[*] By informationally efficient we mean the market on average can correctly order the selections, but for each race this is comparable to randomly splitting the unit interval as explained previously.

[†] This term "ability" is somewhat vague but what is meant by this is the true probability of the horse (and jockey) winning the race.

Chapter 3

Simulating In-Play Horse Odds and A Novel Trading Algorithm

This chapter describes a very common and well-practised hedging strategy known as pairs trading which tries to take advantage of market inefficiencies to produce a profit [47–52]. The strategy will be first explained from a financial markets perspective and then related back to in-play horse racing markets. The approach works on the principle that when highly correlated securities move, they statistically move in the same direction. Some examples of pairs could be two telecom companies, two big corporations that manufacture carbonated sugar and water drinks and two automobile companies. Such pairs could also be components of the same index, which lead to even greater correlations in price. One can describe such a strategy as a statistical arbitrage, since one is relying on historical prices and statistical behaviour to identify the pairs. Once the pairs are identified, then one takes two simultaneous positions in those securities, one short and one long, which is known as a convergence strategy as one hopes for the price of the two securities to converge to the same price in the future. While holding these two positions the trader may exploit market *inefficiencies* which emerge as abnormal prices movement which will result in profit for the trader.

3.1 Market Efficiency Theory

The theory of market efficiency is the idea that markets and their constituent prices are the most efficient means a society can have to distribute aggregated information, famously discussed in [53], but the term efficient market was first coined in [54]. Hayek and Gibson were not aware of an exact mathematical framework for their arguments, but such ideas were phrased in this manner by a French stockbroker, Jules Regnault [55], who noticed that the average deviation of a security's price is directly proportional to the square root of the elapsed time. The more rigorous approach by Louis Bachelier in [56] identified that the actual efficient price of a security behaves as a martingale.

To explain the relationship between martingales and market efficiency we follow a similar line of reasoning to [57]. Consider a financial security has the following price path $\{\dots, S_{t-1}, S_t, S_{t+1}, \dots\}$ the price change is defined as $\Delta S_{t+1} = S_{t+1} - S_t$. The security in question is deemed to have a random payoff X_T at time T in the future. If one could anticipate the prices of the security $\forall t \leq T$ then this would be equal to the expectation of X_T, given all available information up to the present time t and denoted as $\mathcal{F}_t$. This is written mathematically as

$$S_t = \mathbb{E}\left[X_T \mid \mathcal{F}_t\right] \tag{3.1}$$

which is bound by the terminal payoff such that $X_T = S_T$. If we are at the present price S_t, which has the history of prices $\{\dots, S_{t-2}, S_{t-1}, S_t\}$, from Eq. (3.1) the anticipated price would be $S_t = \mathbb{E}\left[S_{t+1} \mid \mathcal{F}_t\right]$ which Implies that $\mathbb{E}\left[\Delta S_{t+1} \mid \mathcal{F}_t\right] = 0$. Thus, if one considers all future payoffs of this security

$$\mathbb{E}\left[\Delta S_{t+1}\Delta S_{t+2}\dots\Delta S_T \mid \mathcal{F}_t\right] = \mathbb{E}\left[\Delta S_{t+1} \mid \mathcal{F}_t\right]\mathbb{E}\left[\Delta S_{t+2} \mid \mathcal{F}_t\right]\dots\mathbb{E}\left[\Delta S_T \mid \mathcal{F}_t\right] = 0, \tag{3.2}$$

therefore all successive price increments are mutually uncorrelated as the price process S_t is a martingale.

The work by Bachelier [56] derived similar results to those found by Einstein, in a different context, five years later [58] and Samuelson sixty-five years later [57]. The significant statistical development towards testing market efficiency as a hypothesis was driven by the work of [59–61] which concluded that prices are martingales and behaviour resembles random walks. For a more thorough review of the development and history of the Efficient Market Hypothesis see [62].

The Efficient Market Hypothesis (EMH) model has been a successful theory, but all theories are falsifiable, and this one is by no means an exception. The EMH has three forms, each differing in the strength of information, albeit public or non-public, on the current price of an asset and how the price adjusts to information flow. The three forms are known as the *strong*, *semi-strong*, and *weak*. The strong form dictates that the current price provides the market with all information including non-public information. The semi-strong relaxes the strength of the previous claim by assuming that the current price only reflects the publicly available information and if new information becomes available the price updates accordingly. The weak form is the least strict in its assumptions of the current and future states of price and only requires the price to be a reflection of the past values (memory). In essence, the relative strength of the EMH is a means of assessing the dynamics of price movements, and if one believes the strong form then price behaves as a martingale (a process where the mean is predictable but not the sample behaviour) and one can not systemically beat the market with any trading strategy. If one rejects this strong form of EMH and believes there are profitable strategies available, then the statistical behaviour will demonstrate patterns that can be exploited for profit. Such patterns could be trends or mean-reversion behaviour. Therefore, if pairs trading is possible, then the behaviour exhibited by the prices is the weak form of EMH.

3.2 Simulating Horse Racing

This section develops a toy model which assumes that markets behave as if it were informational efficient and prices are martingales. In the pursuit of clarity, we will first discuss the notation used in this section: an upper case letter indicates a random variable and lower case letter indicates an observation/sample. A continuous random process with denoted by putting a lowercase d, for example the white noise process is denoted as $dZ_t \sim \mathcal{N}(0,1)$. A discrete random variable is denoted as ΔZ_t.

The model first assumes that all horses start at $X_0 = 0$ and the distance covered relative to the mean distance covered by the competing horses are defined as the following set of independent Itô processes

$$\mathrm{d}X_t^{(i)} = \mu^{(i)}\mathrm{d}t + \sigma^{(i)}\mathrm{d}B_t^{(i)} \tag{3.3}$$

where $t \in (0,1)$ is the time domain, $\mathrm{d}B_t^{(i)} \sim \mathcal{N}(0, \mathrm{d}t)$ is a Wiener process, the drift is sampled from a uniform distribution $\mu^{(i)} \sim \mathcal{U}(a,b)$ such that the real numbers a and b are $a < b$, $\sigma^{(i)}$ is the volatility, which is a positive constant and $i \in \{1, 2, \ldots, N_h\}$ is the index for the N_h horses competing. As a note, the drift parameter $\mu^{(i)}$ is sampled once from the uniform distribution $\mathcal{U}(a,b)$ for each i at time $t = 0$ – which is a similar model as found in Sec. 2.4 – hence, is not too far from the truth. Remembering, there is mapped between a uniform distribution on $(0,1)$ and the exponential distribution, which can be seen by considering the random variables $W \sim \mathcal{U}[0,1]$ and set $Q = -\ln(W)$ then Q is distributed as $F_Q(q) = 1 - F_W(e^{-q})$.

For each horse i and time t we will sample a random variable from a Gaussian density which has the following parameterisation:

$$\left\{\Delta Y_t^{(i)} \sim \mathcal{N}\left(\mu^{(i)}t + x_t^{(i)}, \left(\sigma^{(i)}t'\right)^2\right)\right\}_{m=1,2,\ldots,N_{mc}} \tag{3.4}$$

each independent sample is denoted by the subscript m and performed N_{mc} times, $x_t^{(i)}$ is the present observed value of Eq. (3.3) and $\Delta Y_t^{(i)}$ represents the random displacements for the set of horses at the next time step. Fig. 3.1 gives a visual representation of Eq. (3.4) with $i = 1, 2$ where the density functions for $\Delta Y_t^{(1)}$ and $\Delta Y_t^{(2)}$ are the green and orange distributions respectively. The time variable $t' = 1 - t$ is the complement of time t, decreasing linearly to 0 as t increase to 1 and its use has the effect of reducing the noise as the race comes closer to the finish. The intuitive content of Eq. (3.4) is the centring of a normal distribution

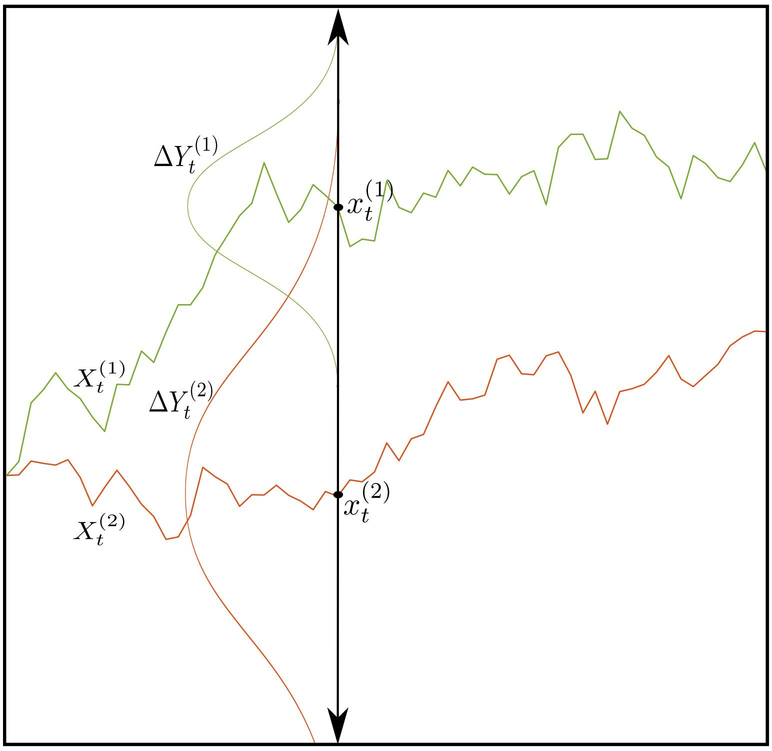


Figure 3.1: A diagram depicting a race with the two horses $X_t^{(1)}$ (green line) and $X_t^{(2)}$ (orange line), both are defined by Eq. (3.3). The current observations of the distanced covered by the two is $x_t^{(1)}$ and $x_t^{(2)}$ each of the distances in the next time step shift by $\Delta Y_t^{(1)}$ and $\Delta Y_t^{(2)}$ respectively, both have densities defined Eq. (3.4). The probability density that horse $X_t^{(2)}$ will be in the pole position in the next time step, $\Delta Y_t^{(1)}$ (green curve) and $\Delta Y_t^{(2)}$ (orange curve) is the density of $\Delta Y_t^{(2)} > \Delta Y_t^{(1)}$.

on $\mu^{(i)}t + x_t^{(i)}$ which is then sampled giving a set of possible future displacements $\Delta Y_t^{(i)}$ that $X_t^{(i)}$ could diffuse to. To implement this model we take a numerical approach using a Monte Carlo model where at each time t the displacements $\Delta Y_t^{(i)}$ are sampled N_{mc} times for each i. Using this numerical method one can calculate the probability that horse i will be in pole position. Consider the points $x_t^{(1)}$ and $x_t^{(2)}$, shown in Fig. 3.1, we generate $\left\{\Delta Y_t^{(1)}\right\}$ and $\left\{\Delta Y_t^{(2)}\right\}$ N_{mc} times each. Counting the normalised frequency that $\Delta Y_t^{(2)} > \Delta Y_t^{(1)}$ gives an approximation of the probability that horse $X_t^{(2)}$ will be in the pole position.

We now explain the numerical approach used to calculate the probability of being in the pole position. The random variable Eq. (3.4) is sampled N_{mc} times for each horse i and time step t. This is interpreted as a Monte Carlo simulation of the future possible states of $Y_t^{(i)}$. Using this Monte Carlo generated sample set $\left\{\Delta Y_t^{(i)}\right\}_{m=1,2,\ldots,N_{mc}}$ which is matrix with dimension $(N_{mc} \times N_h)$ denoted as $\mathbf{Y}_t$. One can find the unconditional probability that horse $X_t^{(i)}$ is in the pole position by finding the frequency in which the elements of i in $\{\mathbf{Y}_t\}_{(m,i)}$ are the maximum value for each m. Mathematically, one defines an indicator matrix which is size $(N_{mc} \times N_h)$ and is denoted as $\mathbf{L}_t$. The elements of $\{\mathbf{L}_t\}_{(m,i)}$ that are set to 1 correspond to the elements of $\{\mathbf{Y}_t\}_{(m,i)}$ that are the maximum value out of the values in the i dimension and the other elements are set to 0. The unconditional probability of horse i at time t being in the pole position is

$$P_t^{(i)} = \frac{\sum_{m=1}^{N_{mc}} \{\mathbf{L}_t\}_{(m,i)}}{N_{mc}} \tag{3.5}$$

where the numerator is the sum of ones and zeros. An example run of this toy model is shown in Fig. 3.2, where the parameters are set as following $N_h = 10$, $N_{mc} = 1 \times 10^4$, $dt \to \Delta t = 1 \times 10^{-2}$, $\sigma^{(i)} = 0.33\ \forall i$ and $\mu^{(i)} \sim \mathcal{U}(0,1)$ sampled for each i once at $t = 0$.

The top plot in Fig. 3.2 is the calculation of Eq. (3.5). One can observe the cointegrating (historically correlated) behaviour as seen in a real in-play market, but because of the way the probability is calculated using Eq. (3.5) there is no over-round. One also observes that the probability of the horse i being in pole position when $t \to 1$ converges to one for the leading horse and zero for the trailing horses.

As a side note it is worth mentioning that one can find an analytical version of Eq. (3.5), that is when the Monte Carlo number $N_{mc} \to \infty$. Consider a race

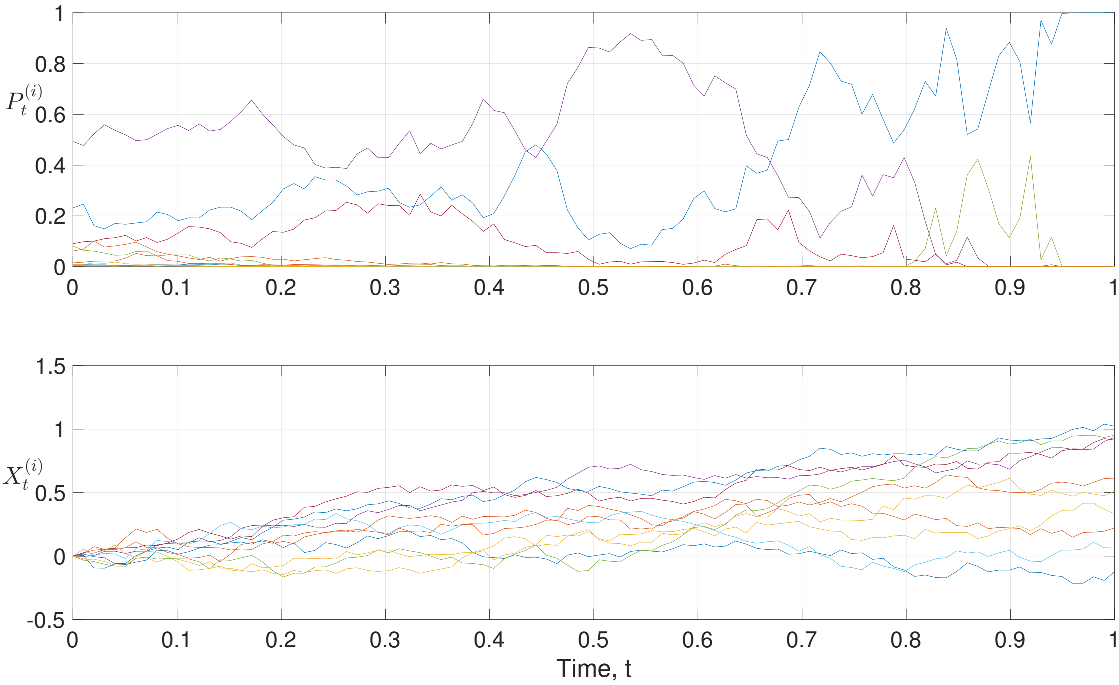


Figure 3.2: Bottom plot is one realisation of $X_t^{(i)}$ which is the numerical integration of the itô process defined in Eq. (3.3). The top plot is the numerical estimation of the unconditional probability that horse i will be in the pole position as defined in Eq. (3.5). The run of this toy model, explained in Sec. 3.2, has the following parameters $N_{mc} = 1 \times 10^4$, $N_h = 10$, $dt \rightarrow \Delta t = 1 \times 10^{-2}$, $\sigma^{(i)} = 0.33\ \forall i$ and $\mu^{(i)} \sim \mathcal{U}\left(0,1\right)$ sampled for each i.

with two horses, as for Eq. (3.4), the performance two density function are

$$\begin{aligned} \Delta Y_t^{(1)} &\sim \mathcal{N}\left(\mu^{(1)} t + x_t^{(1)}, \left(\sigma^{(1)} t'\right)^2\right) \\ \Delta Y_t^{(2)} &\sim \mathcal{N}\left(\mu^{(2)} t + x_t^{(2)}, \left(\sigma^{(2)} t'\right)^2\right) \end{aligned} \tag{3.6}$$

where all variables are same as before. The pdf of $Y = \max\left\{\Delta Y_t^{(1)}, \Delta Y_t^{(2)}\right\}$, denoted as $f(y)$, is found to be $f(x) = f_1(-y) + f_2(-y)$ (see [63]) defining the two pdfs are

$$
\begin{aligned}
f_1(y) &= \frac{1}{\sigma_1}\phi\left(\frac{y+\left(\mu^{(1)}t+x_t^{(1)}\right)}{\sigma^{(1)}t'}\right)\Phi\left(-\frac{y+\left(\mu^{(2)}t+x_t^{(2)}\right)}{\sigma^{(2)}t'}\right)\\
f_2(y) &= \frac{1}{\sigma_2}\phi\left(\frac{y+\left(\mu^{(2)}t+x_t^{(2)}\right)}{\sigma^{(2)}t'}\right)\Phi\left(-\frac{y+\left(\mu^{(1)}t+x_t^{(1)}\right)}{\sigma^{(1)}t'}\right);
\end{aligned}
\tag{3.7}
$$

where the pdf and cdf of the standard normal are denoted respectively as $\phi(.)$ and $\Phi(.)$. The probability measure of Eq. (3.5) is analytically calculated for $P_t^{(2)}$ (odds of horse two taking the pole) as

$$
\mathbb{P}\left[\Delta Y_t^{(1)} < \Delta Y_t^{(2)}\right] = \int_{y_2=-\infty}^{\infty}\int_{y_1=-\infty}^{y_2}\mathcal{N}\left(y_1 \mid M_t^{(1)}, V_t^{(1)}\right)\mathcal{N}\left(y_2 \mid M_t^{(2)}, V_t^{(2)}\right)\mathrm{d}y_1\mathrm{d}y_2
\tag{3.8}
$$

where we have defined the mean and variance respectively as $M_t^{(i)} = \mu^{(i)}t + x_t^{(i)}$ and $V_t^{(i)} = \left(\sigma^{(i)}t'\right)^2$ for $i = 1, 2$.

For a race that consists of more than 2 horses it is little more tricky to analytically calculate Eq. (3.5) but it can be done. The important step is to keep track of the current leading horse by defining the pole position as $j = \arg\max_i\left(x_t^{(i)}\right)$* – such that $\mu^{(i)}t + x_t^{(i)} < \mu^{(j)}t + x_t^{(j)}$ $\forall i$. Once the pole position is defined one can calculate Eq. (3.8) for each i relative to j.

3.3 Pairs Trading

This section explores the statistical arbitrage trading algorithm known as pairs trading. The pairs trading model is explained from a financial market perspective before it is implemented on the real horse racing data that was discussed in Sec. 2.2.

*The arguments of the maxima (abbreviated arg max or argmax) are the points of the domain of some function at which the function values are maximised.

3.3.1 Linear Pricing Models

In order to proceed with the construction of a pairs trading algorithm, one must define the fundamental variables. Let p_i denote the price (today, at time $t = 0$) of asset i and X_i its future payoff (at time $t = 1$). The *(gross) return* on asset i is defined as the ratio between the future payoff and the price now

$$R_i = \frac{X_i}{p_i}. \tag{3.9}$$

The *rate-of-return* on asset i is defined as

$$r_i = R_i - 1 = \frac{X_i - p_i}{p_i} \tag{3.10}$$

and is used as a regressor to identify pairs. The *security market line* (SML) is the graph given by the linear mapping defined by the *capital asset pricing model* (CAPM) [64]. The CAPM is defined as the following: for a given asset's (or portfolio) return one can define a linear relationship such that the i^{th} asset return is

$$r_i = \beta_i(r_m - r_f) + r_f \tag{3.11}$$

where r_i is the expected rate-of-return for the i^{th} asset, β_i is the beta which is a linear measure of the systematic market risk of investment in asset i and r_f is the risk-free rate. The variable r_m is the market rate-of-return defined by

$$r_m = R_m - 1 \tag{3.12}$$

where the subscript m denotes a market or index such as the S&P500 or Euro Stoxx 50. The market beta for $r_i = r_m$ is defined to be $\beta_m = 1$. Assuming a linear correlation between the rate-of-return of the market r_m and the i^{th} asset's rate-of-return, one can find that the betas are related as follows

$$\beta_i = \frac{\mathbb{C}[r_i, r_m]}{\mathbb{V}[r_m]} \tag{3.13}$$

where $\mathbb{C}[.,.]$ and $\mathbb{V}[.]$ are the covariance and variance calculated respectively as

$$\begin{aligned}\mathbb{C}[r_i, r_m] &= \mathbb{E}\left[(r_i - \mathbb{E}[r_i])(r_m - \mathbb{E}[r_m])\right] \quad \text{and} \\ \mathbb{V}[r_m] &= \mathbb{C}[r_m, r_m].\end{aligned} \tag{3.14}$$

β_i is high when market variance is small and r_m is strongly correlated with the i^{th} asset's rate-of-return, and β_i is small when market variance is high and r_m is weakly correlated with the i^{th} asset's rate-of-return.

3.3.2 Market Risk-Neutral

Market beta's are a first order linear approximation of the systematic risk in the investment of an asset from a particular market. If a situation arises by which one has a market beta such that $\beta_i \approx 0$, then this is an indication to the first order that there is no systemic risk and r_i is uncorrelated with the market r_m. Such a situation is known as being *Market Risk-Neutral* and is an objective of an active money manager.

Consider a market participant who has invested in a portfolio consisting of two assets categorised as security 1 and 2. Applying CAPM to each asset suggests that the rates of return are given by

$$\begin{aligned}r_1 &= \beta_1(r_m - r_f) + r_f \\ r_2 &= \beta_2(r_m - r_f) + r_f\end{aligned} \tag{3.15}$$

where the variables are the same as defined in Eq. (3.11). Here one is not giving any indication of how the securities are selected which in practice is important, and more on this is given in Sec. 3.4. One constructs a portfolio with the two assets with the weights π_1 and π_2 (the fraction of the total wealth invested) such that one has the following normalisation $\pi_1 + \pi_2 = 1$. If the total amount of capital to invest in this portfolio is denoted as W_0 (initial wealth), then the wealth invested in each security is $W_1 = \pi_1 W_0$ and $W_2 = \pi_2 W_0$. Given that the total return on such a portfolio is the sum of the fractions invested into each of the assets and then multiplied by their individual returns one can find the beta for this portfolio by applying CAPM, giving

$$r_{12} = \beta_{12}(r_m - r_f) + r_f \tag{3.16}$$

where the portfolio rate-of-return is $r_{12} = \pi_1 r_1 + \pi_2 r_2$ and beta is $\beta_{12} = \pi_1\beta_1 + \pi_2\beta_2$. As before the market neutral strategy is the strategy that yields $\beta_{12} = 0$. To achieve market neutrality, the investor can use particular configurations of the portfolio weights π_1 and π_2, such that $\pi_1\beta_1 + \pi_2\beta_2 = 0$. Having such a market neutral portfolio, where $\beta_{12} = 0$, implies that

$$\begin{pmatrix} \beta_1 & \beta_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{3.17}$$

which has the solution for $\beta_1 \neq \beta_2$

$$\pi_1 = -\frac{\beta_2}{\beta_1 - \beta_2} \text{ and } \pi_2 = \frac{\beta_1}{\beta_1 - \beta_2}. \tag{3.18}$$

Using these two solutions for the relative weights of the investments in the securities 1 and 2, one can calculate the amounts of each security to hold in order to be market neutral. The total number of units to invest in each asset 1 and 2 to be market neutral which is thus

$$N_i = \frac{W_i}{p_i} = \frac{\pi_i W_0}{p_i} \quad \text{for } i = 1, 2 \tag{3.19}$$

where p_i is the current price of asset i and π_i are the weights found in Eq. (3.18). Knowing this result, Eq. (3.19), one has a first order proxy for a market-neutral pairs strategy and if the investor is willing to hold this number of assets they can maintain a first order risk neutral portfolio.

3.3.3 Mean Reversion

Mean reversion is the natural phenomena by which a particular stochastic process will move back (revert) to its mean value. Mean reversion is property that can be exploited to construct a profitable trading strategy. As an example of such a trading strategy, consider a security which has a price that is historically far from the mean price; if this price is known to be a mean reverting process, then the price should return to this mean price. Depending on whether the process is above the mean or below it, a strategy by which one goes short or long in the security, respectively, will statistically make a profit (after exiting the strategy when the reversion is complete). Such a strategy falls into the categories of statistical arbitrage and convergence strategies and these properties are the theoretical basis in the pairs trading strategy construction. To find a mean reversion process, one must consider a set of statistical tests that look to classify if the process in question is mean reverting. The tests usually have a similar goal, which is to compare the statistical differences of a measured process to that of a random walk process. A discrete random walk is mathematically defined as the following process:

$$X_t = X_{t-1} + \varepsilon_t, \tag{3.20}$$

with the initial condition $X_0 = 0$. X_t has independent increments, such increments are Gaussian and are denoted as $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The ε_t term is also known as an idiosyncratic noise, which is a random fluctuation that is endemic to a particular asset's price, for example a stock's price. Such random walks, X_t, are martingales and have zero memory. Consider the sequence $X_1, X_2, X_3, \ldots$ then for any time τ: the sequence satisfies the following: $\mathbb{E}\left[|X_\tau|\right] < \infty$ and $\mathbb{E}\left[X_{\tau+1} \mid X_1, X_2, X_3, \ldots X_\tau\right] = X_\tau$ [17,18,65–72]. The martingale property has the interpretation that the next observed value $X_{\tau+1}$ of the sequence $X_1, X_2, X_3, \ldots X_\tau$ has the conditional expectation of the previous value X_τ, which is equivalent to $\mathbb{E}[X_{\tau+1} - X_\tau \mid X_1, X_2, X_3, \ldots X_\tau] = \mathbb{E}[\varepsilon_t \mid X_1, X_2, X_3, \ldots X_\tau] = 0$. A random walk is the complete opposite to a mean reverting process, as the latter types of process differ by the fact that the mean change in value of the time series is proportional to the current value. The simplest mean reverting process in continuous time is the Ornstein-Uhlenbeck process [73] which is defined as

$$\mathrm{d}X_t = \theta\left(\mu - X_t\right)\mathrm{d}t + \sigma \mathrm{d}W_t \tag{3.21}$$

where θ is the rate of reversion to the mean, μ is the mean value of the process, σ is the volatility of the process and W_t is a Brownian motion or Wiener process

[17, 18, 65–72]. Integrating Eq. (3.21) we obtain the Itô integral

$$X_t = X_0 e^{-\theta t} + \mu \left(1 - e^{-\theta t}\right) + \sigma \int_0^t e^{-\theta(t-s)} \mathrm{d}W_s \tag{3.22}$$

and setting $\mu = 0$ one finds

$$\mathbb{E}\left[X_\tau \mid \{X_t, t \le s\}\right] = X_s e^{-\theta \tau} \quad \forall s \le \tau \tag{3.23}$$

which proves that the Ornstein-Uhlenbeck process, Eq. (3.21), is not a martingale. Note, it is clear that the Ornstein-Uhlenbeck process is not a martingale but it is a Markov process. Therefore, the goal is to find mean reverting price processes to create a statistical arbitrage. Hence, one must use statistical tests for example the Augmented Dickey-Fuller test [74] and in addition a stationarity test, such as a KPSS test [75].

3.3.4 Hurst Exponent

A strictly stationary[*] signal is when the joint probability distribution of the values generated by the process $\{X_t \mid t \ge 0\}$ are invariant under a transformation in time. One would like to test this in order to see if statistical arbitrage strategies can be deployed. A famous test for this is the measure known as the Hurst exponent. This measure explores the repetition of patterns within a signal. Put another way it verifies statistically the memory of a signal. The measure was first described by the hydrologist H.E. Hurst [76] using water level signals for the Nile River. The Hurst Exponent helps us to classify statistically if a signal falls into the categories of random walk, mean-reverting or trending signals. A stochastic process $\{X_t \mid t \ge 0\}$ is called uni-scaling if it has stationary increments and satisfies

$$X_{ct} \sim c^H X_t \tag{3.24}$$

where c is a positive constant and H is the Hurst exponent. If the stochastic process $\{S_t \mid t \ge 0\}$ is a random walk, the variance of changes in the natural

[*] In the future Sec. 5.1.2 we discuss the properties of stationary signals in more detail.

logorithm of this process is calculated as

$$\left\langle \left(\ln(S_{t+h}) - \ln(S_t)\right)^2 \right\rangle \sim h \tag{3.25}$$

where h is a lag, $\langle . \rangle$ is the average over all time points and exponent of h is unity. The common way the Hurst is defined is by modifying Eq. (3.25), such that

$$\left\langle \left(\ln(S_{t+h}) - \ln(S_t)\right)^2 \right\rangle \sim h^{2H} \tag{3.26}$$

and for a random walk one finds $H = 0.5$. Eq. (3.26) quantifies the type of auto-correlation found in signals where such correlations could be classified as long range, short range and uncorrelated. The range of the auto-correlation found can be thought as the memory of past values exhibited by the signal. The fundamental idea is to use Eq. (3.26) to estimate H for a signal and compare this with the Hurst exponent of known signals. The Hurst exponents' values can be summarised as: $H = 0.5$ the series $\ln(S_t)$ is a Brownian motion, $0.5 < H < 1.0$ the series $\ln(S_t)$ is trending (long-memory) and $0 < H < 0.5$ is mean-reverting (anti-persistent). The closer the estimation of the Hurst exponent is to the bounds of $[0, 1]$ the more the signal $\ln(S_t)$ exhibits the behaviour of long-memory ($H = 1$) and anti-persistence ($H = 0$). The Hurst exponent can also be negative and this a phenomena known as *monoscale* [77–79], an effect where signals evolving on small time scales show larger fluctuations than on larger time scales.

Using the last price matched signals discussed in Sec. 2.2, we calculate the Hurst exponent for the entire set of competing horses in all races. Estimating H, one will be able to see the different of types signal auto-correlation behaviour in the in-play horse racing markets. The Fig. (3.3) applies Eq. (3.26) to the last price matched data, where $S_t = LPM_t$ and there are 113999 signals. Fig. (3.3) shows that the Hurst exponent is distributed within the domain $-0.2250 \leq H \leq 0.4480$, suggesting that the signals are mean-reverting and anti-persistent. The left tail of the distribution in Fig. (3.3) becomes negative and this indicates monoscaling in some of the signals. This result indicates that the last price matched signals found within in-play markets are not martingales and thus are not in accordance with the principle of EMH Sec. 3.1. One can also perform a Variance Ratio test [80], which is a standardised test to estimate if a signal is a martingale (a random walk): this test finds that 99.29% of the last price matched signals are not martingales and hence one can further conclude that prices in the in-play horse racing market are not efficient. Therefore, because of this inefficiency one could deploy statistical arbitrage and market making strategies, such as pairs trading, to make a profit.

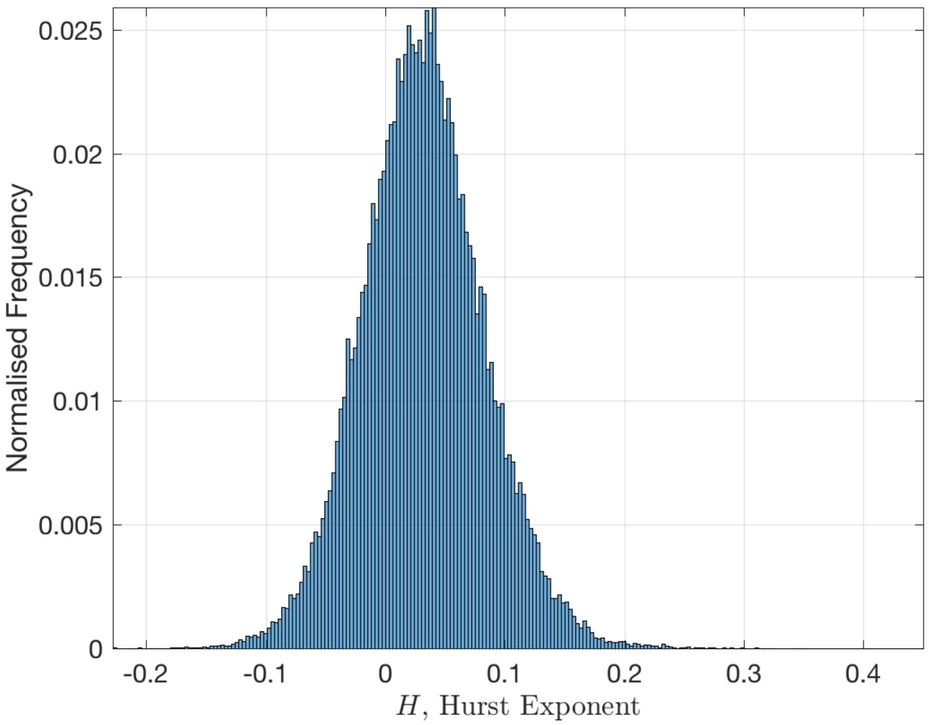


Figure 3.3: The distribution of the Hurst exponent measured on all the $\ln(LPM_t)$ which consists of 113999 race signals. The moments of the distribution are found to be: mean $\langle H \rangle = 0.0316$, standard deviation $\sigma = 0.0511$, skewness $\text{Skew}[H] = 0.1622$, excess kurtosis $\text{Kurt}[H] - 3 = 0.5799$.

3.3.5 Cointegration

This statistical property is at the core of what pairs trading and statistical arbitrage is trying to exploit. When time series are described as being cointegrating pairs, they are exhibiting the property of strong correlation (or anti-correlation); that is on average they move together. To define this in a mathematical context, consider the two time series Y_t and X_t that have stationary covariance under the differencing $\Delta Y_t = (1 - L)^1 Y_t = Y_t - Y_{t-1}$ and $\Delta X_t = (1 - L)^1 X_t = X_t - X_{t-1}$ where one has introduced the lag operator $\Delta^d X_t = (1 - L)^d X_t$ such that

$$L^d X_t = X_{t-d} \, . \tag{3.27}$$

Consider the increment processes ΔY_t and ΔX_t, both of which are stationary. The number of times the difference operator is applied for both processes is $d = 1$, and since both are stationary for $d = 1$, they are referred to as having the *order of integration* 1, which is denoted as $I(1)$ [81]. If the linear combination of Y_t and X_t is found such that the residue process is defined as

$$\varepsilon_t = Y_t - \beta X_t - \alpha \tag{3.28}$$

where α is the interception, $\beta \neq 0$ is a constant parameter to be estimated from observed values of Y_t and X_t. If one finds that the residue process ε_t is stationary for $\beta \neq 0$ then there exists a linear correlation between Y_t and X_t, denoted as $\varepsilon_t \sim I(0)$. The implication of Eq. (3.28) and the stationary property of the residues, $\varepsilon_t \sim I(0)$, is Y_t and X_t are a cointegrating pair because they are correlated. Finding that the residues are not stationary for an order of integration not equal to one is an indication that Y_t and X_t are not a cointegrating pair. The intuition of cointegrating is that in the long-run ($t \to \infty$) the terms in the relationship $Y_t - \beta X_t - \alpha = \varepsilon_t$ will converge to its mean value, such a value is defined as following the time average $\langle \varepsilon_t \rangle = \mu$. Hence, if $\mu = 0$ then in the long-run we have $Y_t = \beta X_t + \alpha$ and this true even if the residues ε_t are autocorrelated.

Cointegration can be used as a means to construct a portfolio that consists of asset prices that by themselves are not mean-reverting but when combined create a portfolio with a value that is mean-reverting. The classic pairs trading strategy is one in which a trader takes a simultaneous long and short position in a financial asset whose prices are cointegrating, in the hope to construct a mean-reverting strategy. Hence to summarise, if one has invested into two equities with prices $P_t^{(1)}$ and $P_t^{(2)}$ forming a linear combination such that $P_t^{(1)} - \beta P_t^{(2)} = \varepsilon_t$ (ignoring α) and finds that $\varepsilon_t \sim I(0)$, then a pair strategy is achievable. The investor in a pairing strategy can also position themselves to be market risk-neutral, Eq. (3.19), ensuring a hedge against linear adverse market movements.

3.4 Pairs Selection in Horse Odds

This section now takes the pairs trading method discussed in Sec. 3.3 and applies it to the in-play horse racing price signals. The algorithm works in the a similar to a historic bookies strategy be which when they match a large risky bets they migrate this risk by making the opposite bet with another bookie (hedging). The signals which are used to detect the pairs are the odds $P_t^{(i)}$, as defined in Eq. (2.1), and this is assumed to be the Last Price Matched $\left(LPM_t^{(i)}\right)$ field found in Table 2.1. One could also use the mid-price to detect the pairs, but we have used the last price matched, in decimal odds form, as a first attempt due to this data field is less sparse than the mid-price. A non-overlapping rolling window of size M, which moves left to right down the $LPM_t^{(i)}$ signal, is used to analyse statistical behaviour that indicates a pair, see Sec. 3.3. The size of the window M is calculated as being the floored value (rounded down to the lowest integer value) which is 10% of the total number of time points. The log-price or log of the decimal odds is defined as $P_t^{(i)} = \ln\left(LPM_t^{(i)}\right)$ where $i \in \{1, 2, \dots, N_h\}$ is the index of the N_h competing horses and $t \in \{1, 2, \dots, M\}$ is the time within each rolling window. The odds $P_t^{(i)}$ are standardised using the following transformation

$$P_t^{\prime(i)} = \frac{P_t^{(i)} - \left\langle P_t^{(i)} \right\rangle}{\sigma^{(i)}}, \tag{3.29}$$

where the sample average is denoted as $\langle . \rangle$ and $\sigma^{(i)}$ is the standard deviation of each of the price processes which are calculated with Eq. (5.2) and Eq. (5.7) respectively. Eq. (3.29) is the z-score transformation and is a statistical measure on a sample set; it gives the relationship between an individual data point (in the sample set) relative to that of the population mean and standard deviation.

The z-score ensures that the prices can be statistically compared and the fluctuations observed are standardised relative to their population's statistics, which is important in the identification of pairs.

To select pairs, we create the following matrix which is the sum of square distances (spreads) within a window

$$\Theta_{ij} = \begin{cases} \sum_{t=1}^{M} \left(P_t^{\prime(i)} - P_t^{\prime(j)} \right)^2, & \text{if } i \neq j \\ 0, & \text{if } i = j. \end{cases} \tag{3.30}$$

The additional index $j \in \{1, 2, \ldots, N_h\}$ is introduced as we are comparing price signals. The matrix Θ_{ij} is a means of estimating the relative fluctuation between each of the odds signals as the race is under way. Hence, a pair would be the relative fluctuations that fluctuate the least. Such an indicator is defined as the vector $\mathbf{p} \in \{1, 2, \ldots, N_h\}^{N_h}$ and $\dim(\mathbf{p}) = (1 \times N_h)$ each element of which is found by the index that gives minimum relative fluctuations. One defines this mathematically in terms of Θ_{ij} as

$$\{\mathbf{p}\}_j = \arg\min_i \left\{\Theta_{ij}\right\}; \text{ such that } i \neq j \tag{3.31}$$

where $\arg\min_i$ returns the index value of $i = 1, 2, \ldots, N_h$ that corresponds to the j that is the minimum value of Θ_{ij}. The intuition here is that one is looking for the price movement j with the least relative divergence with respect to another price which we denote as i^*, hence the price signal would be $P_t^{(i^*)}$. The elements of $\mathbf{p} = \left(i_1^*, i_2^*, \ldots, i_{N_h}^*\right)$ are the first indication that a pair of prices are statistically cointegrating. Using the elements of $\mathbf{p}$ one finds the distance between the normalised prices and the prices at the end of the window, denoted as the set $\left\{P_M^{\prime(i_1^*)}, P_M^{\prime(i_2^*)}, \ldots, P_M^{\prime\left(i_{N_h}^*\right)}\right\}$. So if the solution of Eq. (3.31) is denoted the following vector $\mathbf{p} = \left(i_1^*, i_2^*, \ldots, i_{N_h}^*\right)$ one can find the following vector which is the end of window distances

$$d_j = P_M^{\prime(j)} - P_M^{\prime(i_j^*)} \tag{3.32}$$

for $j = 1, 2, \ldots, N_h$. This vector of distances is also know as the *spread* between the prices at the end of the rolling window at time step M. The Eq. (3.32) and the value of d_j is compared to a control parameter which filters the size of the pairs' spreads to be considered for the pair trade. This ensures that the strategy does not trade a pair unless the spread is greater than a particular level, which is denoted as ϕ. Filtering the spreads of the potential pairs $\mathbf{p}$ such that $d_j > \phi$ one would like to hold a portfolio that is *cost neutral* with the hedging ratio, as seen in Sec. 3.3.2.

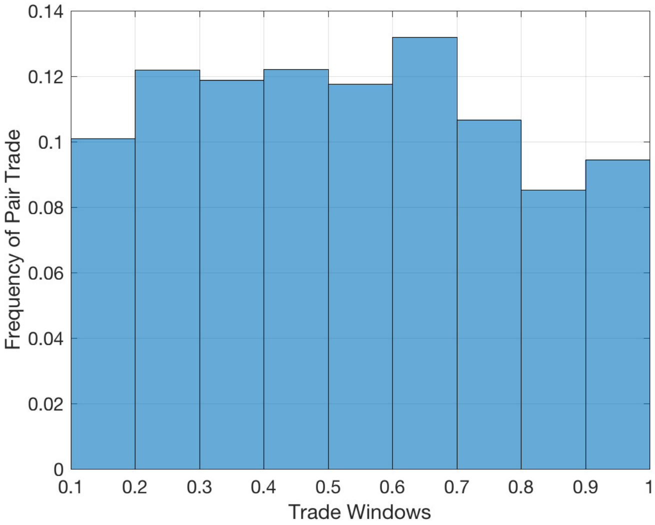


Figure 3.4: The empirical probability of a trade occurring within the time windows indicated on the abscissa. One observes that the implied probability of a pair trade is close to a uniform distribution.

Consider the example where the horse prices $(1, 2)$ are considered to be a pair such that $d_1 > \phi$; one can now define the set of prices within the window for $(1, 2)$ as $x_t = \left\{P_1^{(1)}, P_2^{(1)}, \ldots, P_M^{(1)}\right\}$ and $y_t = \left\{P_1^{(2)}, P_2^{(2)}, \ldots, P_M^{(2)}\right\}$ and fit the following linear model

$$y_t = \beta x_t + \varepsilon_t \tag{3.33}$$

where ε_t are the residuals. The model in Eq. (3.33) is fitted using the ordinary-least-square (OLS) solution which is $\beta = \left(\sum_{t=1}^{M} x_t^2\right)^{-1} \left(\sum_{t=1}^{M} x_t y_t\right)$. The estimation of β is known as the hedging ratio and this gives the relative holding x_t to ensure the strategy is cost neutral. One also uses the hedging ratio β to determine the spread (residuals) between the two signals x_t and y_t

$$\varepsilon_t = y_t - \beta x_t \tag{3.34}$$

and this determines how much to back or lay on the corresponding pair. Using the final residual ε_M one can determine which signals of the pair $(1, 2)$ to back and lay. Defining the parameter γ and the interval $(-\gamma, \gamma)$, if the final residual ε_M is outside this interval:

1. To the negative left one backs y_t with £1 and hedges by laying £β on x_t.

2. To the positive right one lays y_t with £1 and hedges by backing £β on x_t.

One must also check if an open position is on either of the pairs as it should be closed before entering a new one, so as to avoid being over exposed to a particular price signal. We also allow for a 5 time step slippage as there can be a delay when trying to get matched on an order in the Betfair horse racing markets. This process is then repeated to the end of the signal where the positions are recorded and assumed to be matched on £1 wagers and the gains are calculated as Eq. (2.11).

The top plot in Fig. 3.5 shows the accumulated average rate-of-return for the pairs trading strategy. To ensure that there is no bias in picking the races from the 12736 contained in the data set we pick at random 20 races, the pairs strategy is deployed, the rate-of-return is calculated and we repeat this process 1000 times. This average rate-of-return is calculated from the 1000 runs over the 20 races randomly chosen. The top plot in Fig. 3.5 shows that the average rate-of-return positively accumulates and increases ≈ 0.19177 (gradient of the blue line $m = \Delta y/\Delta x$) on average for each race the strategy is deployed on. This result from Fig. 3.5 seems too high (or too good to be true) as by the last race the rate-of-return reaches $\approx 375\%$, this number must be treated with great scepticism. Possible reasons why the rate-of-return increases at such great rate is down to the assumptions of the model: such as when a pair is found the back and lay positions are assumed to be easily matched and it is also assumed that the orders made have no price impact. To measure the strategies reward-to-variability we use the industry standard statistical measure known as the Sharpe ratio [82]. The Sharpe ratio is a measure of signals information to noise, and gives an indication of how much one can gain for a given risk. The bottom plot in Fig. 3.5 are the average Sharpe ratio which is calculated as

$$S_r = \left\langle \frac{R_r}{\sigma_r} \right\rangle \tag{3.35}$$

where the index $r = 1, 2, \ldots 20$ labels the randomly pick race, R_r is the rate-of-return sampled 1000 times for each r then averaged and σ_r is the standard deviation calculated for the 1000 samples. One observes that the average race achieves a Sharpe ratio of $\langle S_r \rangle = 0.3062 \pm 0.0015$ and each S_r calculated in bottom plot of Fig. 3.5 is positive indicating on average that the rate-of-return in positive. The Fig. 3.4 shows the frequency of pair trades deployed in each window, one can see that this is close to a uniform distribution.

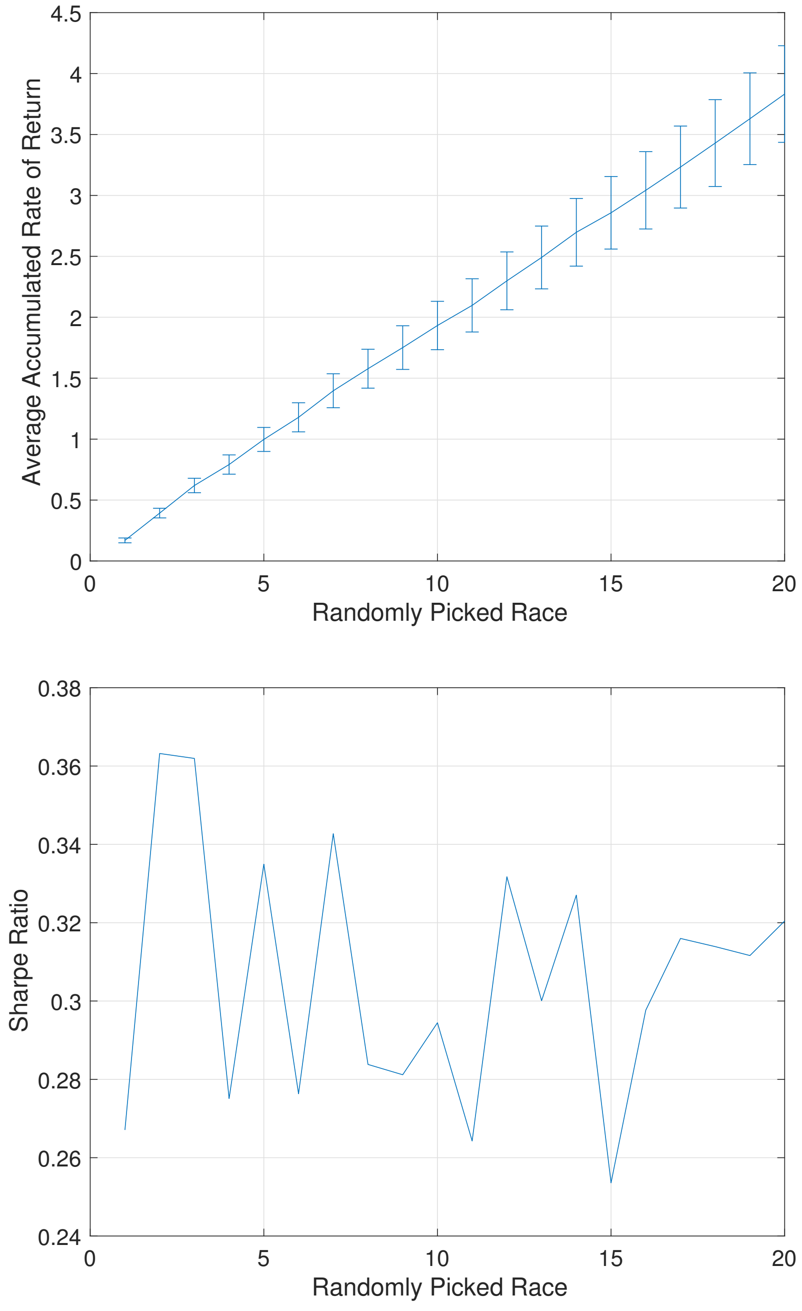


Figure 3.5: Top Plot: The sample average accumulated rate-of-return over 20 randomly picked races sampled 1000 times and averaged over the 1000 samples. Bottom Plot: The sample average Sharpe ratio as calculated in Eq. (3.35) has an average value $\langle S_r \rangle = 0.3062 \pm 0.0015$ and is never negative.

3.5 Summary and Discussion

A novel toy model was described in Sec. 3.2 which assumes the competing horses have a relative position in the race that evolves as a set of Itô processes. These Itô processes are used to calculate the probability numerically with a Monte Carlo strapped on top of each Itô process. One sees from Fig. 3.2 that the odds exhibit a cointegrating behaviour between competing horses. An issue with this model is the static over-round which is always one and it would be interesting to explore the possibility of creating a similar model with a fluctuating over-round as seen in the real data.

In Sec. 3.3 we gave a review of the modern statistical arbitrage trading strategy known as pairs. This trading strategy in Sec. 3.4 is reverse engineered for in-play horse racing data. The algorithm is applied to randomly selected data, and the average rate-of-return is recorded along with the average Sharpe-ratio, Fig. 3.5. These results should be treated with a high degree of scepticism as the last rate-of-return $\approx 370\%$ is suspiciously high and unlikely to be to achieved in reality. The two main reasons why this is so high is down to the assumptions of the model, the first being that backs and lays are always matched and at the front of the queue and the second that the there is no price impact when trading.

Chapter 4

Information Based Finance and Trading

This chapter looks to review the literature which came about from the development of the information-based asset pricing framework. It was first introduced in the following articles [83–85], and will be referred to as the Brody-Hughston-Macrina (BHM) approach. The BHM approach is used as the basis for constructing our own model. Other developments in the area of BHM pricing model have been studied in the following set of literature [86–110]. The BHM methodology initially came about from a split in ideology within the field of *credit risk*. At the time that the BHM approach was developed, the two popular theories that dominated the credit risk literature were, and still are, *structural* and *reduced-form*. Both the structural and reduced-form models are used in the pricing of credit risk products, such as credit derivative swaps (CDS). The two frameworks are discussed and compared in [111] and a more detailed textbook in the field is [112]. To summarise these two approaches, one needs to make assumptions regarding what information is available to the market agents that are pricing credit products; structural models assume that market agents have a *complete* set of information of a firms' financial/credit state and the reduced-form models assume an *incomplete* set of information.

The structural model was first introduced in [113], known as the Merton model, and the first successfully implementation of a structural model was by Kealhofer, McQuown and Vasicek (KMV) in [114]. The basic idea behind structural models is they treat credit events as a class of *first passage time* processes, which are adapted to some predefined stochastic process (filtration). Such processes are calibrated using a company's balance sheet (debt/equity), and defaults are then simulated by observing the frequency with which the calibrated process hits a barrier. Two advantages of structural models: the default event is directly linked to the firm's value and hence its insolvency and the definition of the default time is intuitive. Some drawbacks are the strong assumption that the total value of a firm is a stochastic process that is continuously observable in time and an observed fact from practitioners is that such models underestimate credit spreads for corporate bonds close to maturity [115].

The reduced-form models require fewer assumptions making them more practical than structural models. This approach models the default time as a positive random variable which has a distribution depending on the current state of the economic ecosystem. The reduce-form model was first introduced in [116] and again in [117] with the Jarrow-Turnbull model (for zero-coupon bonds) and popularised in [118] with the Hull and White model (for benchmark bonds). The Hull and White model is one of the most frequently used by practitioners for credit risk pricing. Some advantages of the reduced-form approach are the computational efficiency and pragmatic nature of the model. A drawback of this design is that there are no means of addressing a credit instrument's relationship between the probability of default and its cash-flows.

At the time when the BHM framework was being developed, there was a movement to create a hybrid of the two approaches. This hybrid model would ideally unite the tractability of the reduced-form and the financial events that lead to defaults found in structural models [119, 120].

In the development of the BHM approach, however, there was no similar endeavour taken, and the model falls into the category of reduce-form models.

What makes the BHM model different was a naturally intuitive and economic narrative of explaining an asset's price through its cash-flows.

4.1 Brody-Hughston-Macrina (BHM) Approach

The BHM framework fell in the category of reduced-form modelling and was first applied to the credit instrument known as a *credit derivative.* This is a type of derivative derives its price from an underlying credit instrument, for example, a bond with a fixed deterministic interest rate. A bond is a financial instrument that is purchased with a *face value* is given a maturity. Such bonds may also provide periodic payments known as *coupons* and all such payments to the holder have an additional amount known as the *interest rate* payment. All cash-flows, payment times, maturity and interest rate, are agreed at the time that the bond is issued. Over the lifespan of the bond one can partition the instrument into a series of cash flows (from the buyer's perspective); the initial purchase payment (outflow), the coupon payments received periodically up to maturity and at the maturity (inflow) the final recuperation of the face value (inflow). A product of this form is also referred to as a *debt-obligation* and are considered to be: *default-free* if **all** cash flows are received and in *default* if any of the cash flows are not received. If one assumes that there is a *risk* associated with the due cash-flows, then such an event can be modelled as a random variable measured at predetermined times.

The first assumption of the BHM model is that the bond's history (or reputation) is considered to reveal *partial information* of the cash flow prior to its payment.[*] The second assumption is that the partial information is received temporally and is enshrouded with noise. The true informational content of the cash flow payment is not known: it could be considered to be a binary random variable (payment/non-payment), and this is independently hidden by noise, which is initially assumed to be Gaussian in the BHM model. Before going into the mathematical formalism of this model, one can visualise the second assumption through the following narrative. Consider a television which is currently displaying white

[*] In financial terms this is not too far from the truth as this is the reason bonds are rated by companies such as Moody's and Standard and Poor's.

noise: the viewer at this time can not interpret or perceive any information from this medium as the noise is completely hiding it. Now if the intensity of this obscuring noise is reduced the viewer may be able to interpret that behind the noise is some structure or pattern. The strength is reduced further, and the viewer will see more pattern, for example, the true information could be the results of a group of football matches but since there is still noise they might not be able to make out all the scores or even all the teams playing each other. Reducing the noise completely means the viewer can see all the results are revealed, and therefore all right information is affirmed.

4.1.1 BHM Formalism

Here we outline the mathematics of the BHM model as it was first introduced in [83–85]. Consider a system consisting of a single random future cash flow X_T, such that

$$X_T = \begin{cases} x_1 & \text{payment} \\ x_0 & \text{partial/non-payment} \end{cases} \tag{4.1}$$

where the subscript T denotes the time that the cash-flow is received (or not received). The time domain is defined in the range $0 \leq t \leq T < \infty$ and X_T is $\mathcal{F}_T$-measurable[*]. The values $X_T \in \{x_0, x_1\}$ are categorical variables and can be thought to be equivalent to the binary set $\{0, 1\}$ (or any two member set). This system is modelled by the probability triplet $(\Omega, \mathcal{F}, \mathbb{Q})$ where Ω is the *sample space*, $\mathcal{F}$ is the *σ-field* and $\mathbb{Q}$ the *risk-neutral probability measure* [17, 18, 65–72]. Defining the probability triplet is not enough if one wishes to model the flow of partial information regarding the true value of X_T at times $0 < t < T$. To model this, partial information regarding the future cash flow at times before T, namely the process $\{\mathcal{F}_t\}_{0 \leq t < T < \infty}$ is defined and is known in the literature as the *market filtration* or *natural filtration*. A natural filtration is the σ-field generated by the historical dynamics of a process at each time; consider the process $\{S_t\}_{0 \leq t < T}$ then the natural filtration is

$$\mathcal{F}_t = \mathcal{F}_t^S \triangleq \sigma\left(\{S_u \,|\, 0 \leq u \leq t\}\right), \ \forall t \in [0, T] \ . \tag{4.2}$$

[*] X_T is not $\mathcal{F}_t$-measurable hence the value of X_T is unknown till it is observed at time T.

In the standard setup, one would assume all price processes to be adapted to a pre-specified market filtration, for example the Black-Scholes-Merton model assumes all price processes denoted as $\{S_t\}_{0 \le t < T < \infty}$ are all adapted to the σ-field generated by a geometric Brownian motion.[†] The important difference in the BHM approach, as compared to the standard pricing setup, is that one is seeking to find the explicit means to model the flow of information by giving more structure to the market filtration $\{\mathcal{F}_t\}$.

Assuming that the system is arbitrage free implies the existence of *physical probability measure* $\mathbb{P}$ that is equivalent to the risk-neutral measure $\mathbb{P} \sim \mathbb{Q}$. Therefore, this assumption allows the use of the *fundamental theorem of asset pricing* [17, 18, 65–72] which ensures that a pricing kernel exists. The pricing kernel is thus

$$S_t = D_{tT} \mathbb{E}^{\mathbb{Q}} \left[X_T \mid \mathcal{F}_t \right] \tag{4.3}$$

where D_{tT} is the discounted default-free bond value which is calculated as

$$D_{tT} = e^{-\int_t^T r(u)\,\mathrm{d}u} \tag{4.4}$$

where the deterministic spot rate is denoted as $r(t)$. D_{tT} is a function which is bound in the half closed interval $D_{tT} \in (0,1]$ and is monotonically decreasing to zero as $\lim_{T\to\infty} D_{tT} \to 0$ and has the time derivative $\frac{dD_{tT}}{dt} \le 0$ almost everywhere.

At the terminal time $t = T$ the random variable X_T will be revealed with the probabilities $\mathbb{Q}[X_T = x_0] = p_0$ and $\mathbb{Q}[X_T = x_1] = p_1$ such that: $p_0, p_1 \in (0,1)$, $p_0 + p_1 = 1$ and $x_0 < x_1$. Given the pricing kernel Eq. (4.3) the initial price at $t = 0$ is found to be

$$S_0 = D_{0T} \mathbb{E}^{\mathbb{Q}} \left[X_T \mid \mathcal{F}_0 \right] = D_{0T} \mathbb{E}^{\mathbb{Q}} \left[X_T \right], \tag{4.5}$$

which results in the initial price $S_0 = D_{0T}(p_0 x_0 + p_1 x_1)$. The *a priori* probabilities p_0 and p_1 can be solved in terms of the future cash flows x_0, x_1, S_0 and D_{0T} [83–85], and thus can be statistically calibrated to real market data.

[†] Mathematically such a filtration would be donated as the following process $\{\mathcal{F}_t\} = \sigma(\{S_t\})$ where $0 \le t < T < \infty$ and $\sigma(.)$ is the sigma algebra generator.

4.1.2 True Information

The cash flow X_T is defined as before in Eq. (4.1) and is $\mathcal{F}_T$-measurable. The true information in the system regarding the final value of X_T is received in addition to noise (partial information), before it is finally revealed at $t = T$. This is modelled in the same manner as the X-factor analysis described in [83–110].

We define the constant σ (not to be confused with a σ-algebra) as the rate of information revelation: it represents the intensity in the trueness of information, that is

- High values of σ corresponds to a cash flow values that are revealed quickly.
- Small values of σ corresponds to a cash flow values that are revealed slowly.

In the BHM model as formalised in [83–110] in the literature it is assumed that the true information is revealed linearly as $\sigma X_T t$ (there does not seem to be any economic or financial justification for this other than mathematical convenience). The rate of information is $[\sigma] = [\text{price}]^{-1}[\text{time}]^{-1/2}$ [83,84], which can be seen later on from Eq. (4.23) to ensure that the exponent is dimensionless when pricing; therefore it is apparent why one calls this a rate. As a final note this form of the true information has no explanation stemming from reality and the case maybe that it is a function such that $\sigma(t) \neq \sigma t$.

4.1.3 Brownian Bridge

Following the original framework set out in the initial BHM model [83–85], one parameterises the noise component of the information signal using a Brownian bridge process. The Brownian bridge [17, 18, 65–72] has the following properties:

Given a standard Brownian bridge process $\{\beta_{tT}\}_{0\leq t\leq T}$, which exists in the time interval $t \in [0, T]$, and is independent of the cash-flow under the risk-neutral measure $\mathbb{Q}$ (this independence is denoted as $\beta_{tT} \perp\!\!\!\perp^{\mathbb{Q}} X_T$). The standard Brownian Bridge is distributed as $\beta_{tT} \sim \mathcal{N}(0, t(T-t)/T)$ and is defined as

$$\beta_{tT} \triangleq B_t - \frac{t}{T} B_T \quad 0 \leq t \leq T \tag{4.6}$$

with the start and end values set to $\beta_{0T} = \beta_{TT} = 0$. The process $\{B_t\}_{0 \le t \le T}$ is an $\{\mathcal{F}_t\}$–adapted standard Brownian motion and B_T is a random variable which is $\mathcal{F}_T$–measurable. The standard Brownian bridge process as defined in Eq. (4.6) can be numerically simulated and sample realisations are shown in Fig. 4.1. The Brownian bridge process $\{\beta_{tT}\}_{0 \le t \le T}$ by definition is **not** an $\{\mathcal{F}_t\}$-adapted process. This property has the important interpretation that market participants in the time domain $0 < t < T$ cannot distinguish between noise and true information until a cash flow is paid at $t = T$.

The expectation of the Brownian bridge, defined in Eq. (4.6), gives $\mathbb{E}[\beta_{tT}] = 0$ and the signal's auto-covariance is trivially found to be

$$\mathbb{C}[\beta_{sT}, \beta_{tT}] = \mathbb{E}[\beta_{sT}\beta_{tT}] = \frac{s(T-t)}{T} \qquad 0 \le s \le t \le T. \tag{4.7}$$

The distribution of the Brownian bridge's increments is given by

$$\mathrm{d}\beta_{tT} = \mathrm{d}B_t - \frac{B_T}{T}\mathrm{d}t \tag{4.8}$$

where the increments are normally distributed as $\mathrm{d}B_t \sim \mathcal{N}(0, \mathrm{d}t)$ and the random variable B_T is $\mathcal{F}_T$-measurable. One can also write Eq. (4.8) equivalently as

$$\mathrm{d}\beta_{tT} = -\frac{\beta_{tT}}{T-t}\mathrm{d}t + \mathrm{d}B_t \tag{4.9}$$

where $\beta_{0T} = 0$, which is a well known result that can found in textbooks such as [121]. Eq. (4.9) has the following integral solution

$$\beta_{tT} = \begin{cases} (T-t)\int_0^t \frac{\mathrm{d}B_u}{T-u} & 0 \le t < T \\ 0 & t = T \end{cases} \tag{4.10}$$

which is equivalent to Eq. (4.6) and allows one to dynamically create such a process with a computer.

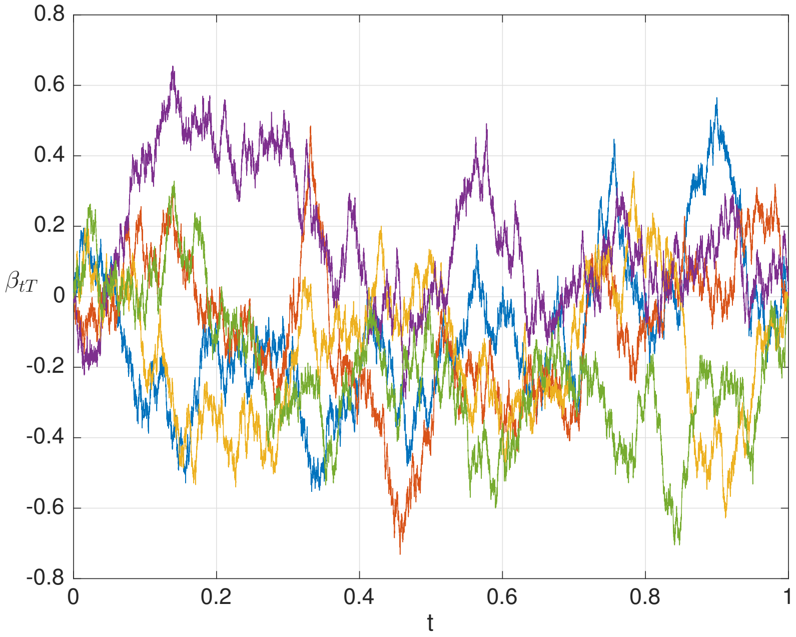


Figure 4.1: Five independent standard Brownian bridge processes as defined in Eq. (4.6) with the time step $\Delta t = 1 \times 10^{-5}$ and $T = 1$.

4.1.4 Information Process

Combining the true information term and Brownian bridge, one derives the information process of the BHM model. The cash-flow X_T is revealed at time T. This value is inaccessible to any market participants and is not $\mathcal{F}_t$-measurable for $t < T$ but it is in fact $\mathcal{F}_T$-measurable. The filtration $\mathcal{F}_t$ is assumed to represent the time dependent flow of noisy information in regard to the value of X_T. The BHM approach theorises that the following process characterises such a flow of information

$$\xi_t = \sigma X_T t + \beta_{tT} \tag{4.11}$$

where the noise is generated by the term $\{\beta_{tT}\}_{0 \leq t \leq T}$ which is a standard *Brownian bridge* process (discussed in Sec. 4.1.3) and the signal representing the flow of *true information* is $\sigma X_T t$ (discussed in Sec. 4.1.2). The true information term is linear in time meaning revelation of true information regarding the future cash-flow X_T grows at a linear rate σ which is a positive constant. The process defined in Eq. (4.11) is referred to in the literature as an *information process* [83–110].

An information process is assumed to be accessible to the market at times $t < T$ and it is implied that $\{\xi_t\}_{0\leq t<T}$ is $\{\mathcal{F}_t\}$-adapted. The filtration generated by the information process is $\left\{\mathcal{F}_t^{\xi}\right\} = \sigma\left(\{\xi_t\}_{0\leq s<t}\right)$ and this is a sub-σ-algebra of the natural filtration, Eq. (4.2), which is denoted as $\left\{\mathcal{F}_t^{\xi}\right\} \subseteq \{\mathcal{F}_t\}$. The crucial step in BHM approach is to make the assumption that the natural filtration $\{\mathcal{F}_t\}$ is adapted to $\sigma-$algebra generated by the information process $\{\xi_t\}$, that is $\{\mathcal{F}_t\} \triangleq \left\{\mathcal{F}_t^{\xi}\right\}$. Hence the price process Eq. (4.3) is

$$S_t = D_{tT}\mathbb{E}^{\mathbb{Q}}\left[X_T \mid \mathcal{F}_t^{\xi}\right] \tag{4.12}$$

and since the information process in Eq. (4.11) is a Markov process one can rewrite this as

$$S_t = D_{tT}\mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t\right] \tag{4.13}$$

and this form of the pricing kernel will be discussed in more detail in the proceeding section, Sec. 4.2.1.

We now discuss the Markov property and why one can rewrite Eq. (4.12) as Eq. (4.13). Consider the information process defined in Eq. (4.11), if such a process is Markov then we have

$$\mathbb{Q}\left[\xi_t \leq y \mid \xi_u\right] = \mathbb{Q}\left[\xi_t \leq y \mid \mathcal{F}_u^{\xi}\right], \quad \forall y \in \mathbb{R} \tag{4.14}$$

where all times u and t are $0 \leq u \leq t \leq T$. The Markov property (shown Eq. (4.14)) is a characteristic of a subset of stochastic processes where the conditional probability distribution only depends on the present state – when the process is conditioned on both past and present states the process (here denoted as $\mathbb{Q}\left[. \mid \xi_u\right]$). To prove Eq. (4.14) consider a sequence of increasing times (but decreasing index) $0 < u_n < u_{n-1} < \cdots < u_2 < u_1 < t \leq T$. Using this time sequence one can define the following sequence of random variables

$$\kappa_i \triangleq \frac{\beta_{u_i T}}{u_i} - \frac{\beta_{u_{i+1} T}}{u_{i+1}} = \frac{\xi_{u_i}}{u_i} - \frac{\xi_{u_{i+1}}}{u_{i+1}} \tag{4.15}$$

where $i = 1, 2, \ldots, n-1$ and the covariance between κ_i and β_{tT} is zero. This zero covariance is shown if one takes the following expectation

$$\mathbb{E}^{\mathbb{Q}}\left[\beta_{tT}\kappa_i\right] = \frac{u_i(T-t)}{u_iT} - \frac{u_{i+1}(T-t)}{u_{i+1}T} = 0 \quad \forall i \tag{4.16}$$

hence $\kappa_1, \kappa_2, \ldots, \kappa_{n-1}$ are independent. Applying the conditional probability in Eq. (4.14)) to the sequence of Gaussian random variables $\xi_{u_1}, \xi_{u_2}, \ldots, \xi_{u_n}$ gives

$$\begin{aligned}\mathbb{Q}\left[\xi_t \leq y \mid \xi_{u_1}, \xi_{u_2}, \ldots, \xi_{u_n}\right] &= \mathbb{Q}\left[\xi_t \leq y \mid \xi_{u_1}, \frac{\xi_{u_1}}{s_1} - \frac{\xi_{u_2}}{u_2}, \frac{\xi_{u_2}}{u_2} - \frac{\xi_{u_3}}{s_3} -, \ldots, \frac{\xi_{u_{n-1}}}{u_{n-1}} - \frac{\xi_{u_n}}{u_n}\right] \\ &= \mathbb{Q}\left[\xi_t \leq y \mid \xi_{u_1}, \frac{\beta_{u_1T}}{s_1} - \frac{\beta_{u_2T}}{u_2}, \frac{\beta_{u_2T}}{u_2} - \frac{\beta_{u_3T}}{s_3} -, \ldots, \frac{\beta_{u_{n-1}T}}{u_{n-1}} - \frac{\beta_{u_nT}}{u_n}\right] \\ &= \mathbb{Q}\left[\xi_t \leq y \mid \xi_{u_1}, \kappa_1, \kappa_2, \ldots, \kappa_{n-1}\right]\end{aligned} \tag{4.17}$$

and using the definition of conditional probability on the left hand side

$$\begin{aligned}\mathbb{Q}\left[\xi_t \leq y \mid \xi_{u_1}, \kappa_1, \kappa_2, \ldots, \kappa_{n-1}\right] &= \frac{\mathbb{Q}\left(\xi_t \leq y, \xi_{u_1} \leq y_0, \kappa_1 \leq y_1, \kappa_2 \leq y_2, \ldots, \kappa_{n-1} \leq y_{n-1}\right)}{\mathbb{Q}\left(\xi_{u_1} \leq y_0, \kappa_1 \leq y_1, \kappa_2 \leq y_2 \ldots, \kappa_{n-1} \leq y_{n-1}\right)} \\ &= \frac{\mathbb{Q}\left(\xi_t \leq y, \xi_{u_1} \leq y_0\right)\mathbb{Q}\left(\kappa_1 \leq y_1\right)\mathbb{Q}\left(\kappa_2 \leq y_2\right), \ldots, \mathbb{Q}\left(\kappa_{n-1} \leq y_{n-1}\right)}{\mathbb{Q}\left(\xi_{u_1} \leq y_0\right)\mathbb{Q}\left(\kappa_1 \leq y_1\right)\mathbb{Q}\left(\kappa_2 \leq y_2\right), \ldots, \mathbb{Q}\left(\kappa_{n-1} \leq y_{n-1}\right)} \\ &= \mathbb{Q}\left(\xi_t \leq y \mid \xi_{u_1}\right)\end{aligned} \tag{4.18}$$

which proves that the information process $\{\xi_t\}_{t>0}$ is a Markov process.

4.2 BHM Pricing Review

This section explores the pricing of a binary bond in the information process framework which uses Eq. (4.13). A later goal of this chapter is to explore and construct a model which gives a trading mechanism to agents, which was not fully investigated in BHM [83–110]. To a achieve this one must first determine the pricing of a binary class of asset (which is discussed in Table 4.2.1) and then find the dynamics of the associated price process (which is discussed in Sec. 4.2.2). The dynamics of the price processes are implemented for each of the agents in a synthetic market where their pricing perception change when they trade with each other; this is allowed to evolve in the time interval $t \in (0, T)$. As the agents' price processes materialise through the time interval, disagreements are brought about via the agents' independent Brownian bridge processes. The Brownian bridge processes can be thought of as the agents' reaction to noisiness of the partial information (be it true or false) concerning the price of the future payoff of the asset.

From a financial market's point-of-view prices are submitted to a *clearing house* which is the mechanism by which buyers and sellers exchange financial instruments for capital. Clearing houses are an important mechanism of aiding the liquidity of financial markets without revealing the parties that are trading. It has also been known that a clearing house can also suggest a price if buyers and seller are within a threshold price of each other. Within the clearing house system one can place orders for a particular asset. When disagreements between buying and selling prices of these orders arise one has an *order book* [122–125]. The difference between the prices of the most competitive buy and sell order is called the *spread*. This trading mechanism will be used as a away of developing the BHM approach by incorporating trading. Trading gives rise to a *market price* which evolves in the time interval $t \in (0, T)$ but disagreements in price are driven by the agents' asymmetric information and will be investigated in this section.

4.2.1 BHM Pricing of a Simple Bond

Consider an asset which consists of a single cash flow that is revealed at $t = T$, where $T \in \mathbb{R}_{>0}$. The receiving of the cash flow can be categorised as a random variable X_T which takes a value in the range $[0, 1]$. To price such an asset in the time interval $t \in (0, T)$, we use Eq. (4.13), which is the following conditional expectation

$$S_t = \mathbb{E}^{\mathbb{Q}}[X_T \mid \xi_t] \tag{4.19}$$

where $\mathbb{Q}$ is the risk-neutral probability distribution and $\{\xi_t\}_{t>0}$ the information process as described by Eq. (4.11). The price process $\{S_t\}_{t\leq 0}$ generated by Eq. (4.19) is a *martingale*[*]. Note the removal, for simplicity, of any contribution

[*] A stochastic process $\{Y_t\}_{t\geq 0} \in \mathbb{R}$ is called a martingale with respect to process $\{X_t\}_{t\geq 0} \in \mathbb{R}$ that is defined on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{Q})$ where $\{\mathcal{F}_t\}_{t\geq 0} = \sigma(\{X_t\}_{t\geq 0})$ $\forall t \geq 0$, when two properties hold: (i) $\mathbb{E}[|Y_t|] < \infty$ and (ii) $\mathbb{E}[Y_t \mid \mathcal{F}_s] = Y_s \quad \forall s \leq t$.

of discounted future cash flow by setting the risk-free rate $r = 0$ and hence the discount factor is unity $D_{tT} = 1 \; \forall t \in [0, T]$. The conditional expectation Eq. (4.19) is found by the following integral

$$S_t = \int_0^1 x \, \mathrm{d}\mathbb{Q}(X_T \leq x \mid \xi_t) \; , \tag{4.20}$$

but to evaluate this one needs to define the conditional posterior probability density function in term of $\mathbb{Q}(.)$. The posterior conditional probability density function of the continuous random variable X_T for given observation of the information $\xi_t = y$ is defined as the following differential

$$f_{X_T}(x \mid \xi_t = y) \triangleq \frac{\mathrm{d}}{\mathrm{d}x} \mathbb{Q}(X_T \leq x \mid \xi_t = y) \; , \tag{4.21}$$

With the application of Bayes' theorem such a conditional probability density can be written as

$$f_{X_T}(x \mid \xi_t = y) = \frac{f_{X_T}(x) f_{\xi_t}(y \mid X_T = x)}{\int_0^1 f(z) f_{\xi_t}(y \mid X_T = z) \, \mathrm{d}z} \; , \tag{4.22}$$

where $f_{X_T}(x)$ is the prior belief of the cash-flow X_T and $f_{\xi_t}(y \mid X_T = x)$ is the likelihood function of observing information y given that the cash-flow is $X_T = x$. The information process, Eq. (4.11), is distributed as a Gaussian denoted as $\xi_t \sim \mathcal{N}(\sigma x t, t(T-t)/T)$. Substituting into Eq. (4.22), one can expressed the conditional posterior density of the cash flow as

$$f_{X_T}(x \mid \xi_t = y) = \frac{f_{X_T}(x) \exp\left(\frac{T}{T-t} \sigma x (y - \frac{1}{2} \sigma x t)\right)}{\int_0^1 f_{X_T}(z) \exp\left(\frac{T}{T-t} \sigma z (y - \frac{1}{2} \sigma z t)\right) \mathrm{d}z} . \tag{4.23}$$

where the information rate parameter is $\sigma \in \mathbb{R}_{>0}$. To streamline the notation for the probability densities, one can define the prior as $f_{X_T}(x) \triangleq f(x)$ and posterior as $f_{X_T}(x \mid \xi_t = y) \triangleq \phi_t(x)$, such that

$$\phi_t(x) = \frac{f(x) \exp\left(\frac{T}{T-t} \sigma x (y - \frac{1}{2} \sigma x t)\right)}{\int_0^1 f(z) \exp\left(\frac{T}{T-t} \sigma z (y - \frac{1}{2} \sigma z t)\right) \mathrm{d}z} . \tag{4.24}$$

This conditional probability density $\phi_t(x)$ is central to how one prices in the BHM framework. The posterior $\phi_t(x)$ is updated when there is an update in the likelihood function $f_{\xi_t}(y \mid X_T = x)$ as the two are directly proportional $\phi_t(x) \propto f_{\xi_t}(y \mid X_T = x)$. In addition, the likelihood function is dependent on the information process $\{\xi_t\}$, meaning changes in $\{\xi_t\}$ update the posterior. Using $\phi_t(x)$, Eq. (4.19) and Eq. (4.20) the price is then calculated as the following integration

$$S_t = \int_0^1 x\phi_t(x)\,\mathrm{d}x \to (0,1), \tag{4.25}$$

where one notices that because the likelihood function is dependent on the information process $\{\xi_t\}$, then the price is function of $S_t \equiv S(\xi_t, t)$. Formally price is a function with the following mapping $S : \mathbb{R} \times [0,T) \to (0,1)$.

4.2.2 Dynamic Analysis of a Simple Bond

This section looks at the derivation of the dynamics of the price process $\{S_t\}$ found in Eq. (4.25), which takes a similar method to [83, 86], but setting the interest rate $r = 0$. Since the price can be written as a function of the information process ξ_t and the time t, one can derive the underlying dynamics of the price process $\mathrm{d}S_t$. The increment of price $\mathrm{d}S_t$ is found by differentiating $S(\xi_t, t)$ with Itô's lemma

$$\mathrm{d}S_t = \left(\frac{\partial S(\xi_t,t)}{\partial t} + \frac{1}{2}\frac{\partial^2 S(\xi_t,t)}{\partial \xi_t^2}\right)\mathrm{d}t + \frac{\partial S(\xi_t,t)}{\partial \xi_t}\mathrm{d}\xi_t\ , \tag{4.26}$$

but one first needs to show that the information process is a Lévy process in order to apply Itô's lemma. The dynamics of the information process $\mathrm{d}\xi_t$ is found by differentiating Eq. (4.11) and using Eq. (4.9) as the definition of the Brownian bridge. One finds

$$\begin{aligned}\mathrm{d}\xi_t &= \mathrm{d}\left[\sigma X_T t + \beta_{tT}\right] \\ &= \sigma X_T \mathrm{d}t - \frac{\beta_{tT}}{T-t}\mathrm{d}t + \mathrm{d}B_t \\ &= \frac{1}{T-t}\left(\sigma X_T T - \xi_t\right)\mathrm{d}t + \mathrm{d}B_t\end{aligned} \tag{4.27}$$

which is *semimartingale*[*], and the quadratic variation of the information process is shown to be

$$\mathbb{E}[\mathrm{d}\xi_t^2] = \mathbb{E}\left[\left(\left(\sigma X_T - \tfrac{B_T}{T}\right)\mathrm{d}t + \mathrm{d}B_t\right)^2\right] = \mathrm{d}t. \tag{4.28}$$

Since it has been shown from Eq. (4.27) that ξ_t is a semimartingale and Eq. (4.28) has finite quadratic variation, then $\{\xi_t\}$ is a Lévy process and therefore one can apply Itô's lemma to $S(\xi_t, t)$ as in Eq. (4.26). The three partial derivative terms in Eq. (4.26) are found to be

$$\begin{aligned}
\frac{\partial S(\xi_t,t)}{\partial t} &= \int_0^1 \mathrm{d}x\, x \frac{\partial}{\partial t}\phi_t(x) \\
&= \int_0^1 \mathrm{d}x\, x\phi_t(x)\frac{\sigma T}{(T-t)^2}\left(xy - \frac{1}{2}\sigma x^2 T - yS(\xi_t,t) + \frac{1}{2}\sigma T\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t = y\right]\right) \\
\frac{\partial S(\xi_t,t)}{\partial \xi_t} &= \int_0^1 \mathrm{d}x\, x \frac{\partial}{\partial y}\phi_t(x) \\
&= \int_0^1 \mathrm{d}x\, x\phi_t(x)\frac{\sigma T}{T-t}\left(x - S_t\right) \\
\frac{\partial^2 S(\xi_t,t)}{\partial \xi_t^2} &= \int_0^1 \mathrm{d}x\, x \frac{\partial^2}{\partial y^2}\phi_t(x) \\
&= \int_0^1 \mathrm{d}x\, x\phi_t(x)\frac{\sigma^2 T^2}{(T-t)^2}\left((x - S_t)^2 - \int_0^1 \mathrm{d}x\phi_t(x)\left(x - \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t = y\right]\right)^2\right)
\end{aligned} \tag{4.29}$$

which leads to the following dynamic expression for the price increment process

$$\mathrm{d}S_t = \frac{\sigma T}{T-t}V_t\left(\frac{1}{T-t}(\xi_t - \sigma T S_t)\mathrm{d}t + \mathrm{d}\xi_t\right) \tag{4.30}$$

where the conditional time dependent variance of the cash flow X_T is denoted as V_t. The conditional variance of the cash flow is defined as $V_t \triangleq \mathbb{V}[X_T \mid \xi_t = y]$ and

[*] A process $\{X_t\}_{t\geq 0} \in \mathbb{R}$ defined on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{Q})$ is called a semimartingale if it can be decomposed as $X_t = M_t + A_t$ where M_t (in Eq. (4.27) term $\int_0^t \mathrm{d}B_s$) is a local martingale and A_t (in Eq. (4.27) term $\int_0^t \frac{1}{T-s}\left(\sigma X_T T - \xi_s\right)\mathrm{d}t$) is a RCLL ("right continuous with left limits" adapted process of locally bounded variation).

explicitly calculated as

$$\begin{aligned}V_t &= \int_0^1 \mathrm{d}x\phi_t(x)\left(x - \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t = y\right]\right)^2 \\ &= \int_0^1 x^2\phi_t(x)\mathrm{d}x - \left(\int_0^1 x\phi_t(x)\mathrm{d}x\right)^2\end{aligned}$$

which written in expectation terms is

$$V_t = \mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right] - \left(\mathbb{E}^{\mathbb{Q}}[X_T \mid \xi_t]\right)^2. \tag{4.31}$$

The conditional variance can be described as a process $\{V_t\}_{0\leq t\leq T}$ such that $\mathbb{E}^{\mathbb{Q}}\left[|V_T|\ \right] < \infty$ and $\mathbb{E}^{\mathbb{Q}}\left[V_T \mid \xi_t\right] \leq V_t\ \forall t \leq T$ implying that V_t is a *supermartingale*. The history of V_t tends to be bound from below (to observe this supermartingale property for a binary bond see Fig. 4.2 and Fig. 4.3). One can see that V_t is a *supermartingale* from Eq. (4.31) because Eq. (4.3) is a *martingale* using Jensens' inequality, the square of a martingale (the term on the right of the minus sign in Eq. (4.31)) is a *submartingale* and $\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right]$ is martingale implying that V_t is a supermartingale because the difference between a martingale and submartingale gives a supermartingale [126]. The conditional variance V_t being a supermartingale indicates that the average uncertainty of the terminal payoff X_T tends to decrease in time ($t \to T$).

One can show that Eq. (4.30) can be written as the following dynamic process

$$\mathrm{d}S_t = \frac{\sigma T}{T-t}V_t\mathrm{d}\tilde{B}_t \tag{4.32}$$

where $\left\{\tilde{B}_t\right\}_{0\leq t<T}$ is a standard Brownian motion and is independent of the standard Brownian motion $\{B_t\}_{0\leq t<T}$, which drives the Brownian bridge process Eq. (4.6) and hence the information process Eq. (4.11).

To prove Eq. (4.32) one must show that

$$\mathrm{d}\tilde{B}_t = \left(\frac{1}{T-t}(\xi_t - \sigma T S_t)\mathrm{d}t + \mathrm{d}\xi_t\right) \tag{4.33}$$

and therefore prove that $\{\tilde{B}_t\}_{0\leq t<T}$ and $\{B_t\}_{0\leq t<T}$ are both:

(i) Independent Brownian motions

(ii) Martingales

First to prove point (i), consider the increments in Eq. (4.33) such that

$$\tilde{B}_s - \tilde{B}_t = \xi_s - \xi_t + \int_t^s \frac{1}{T-u}(\xi_u - \sigma T S_u)\mathrm{d}u \tag{4.34}$$

where $0 \leq t \leq s < T$ and the time step $s - t$ is small. The increments in the Brownian bridge process, defined in Eq. (4.9), can be written as

$$\beta_{sT} - \beta_{tT} = -\int_t^s \frac{\beta_{uT}}{T-u}\mathrm{d}u + B_s - B_t \tag{4.35}$$

and Eq. (4.34) becomes

$$\tilde{B}_s - \tilde{B}_t = \sigma(s-t)X_T - \int_t^s \frac{\beta_{uT}}{T-u}\mathrm{d}u + B_s - B_t + \int_t^s \frac{1}{T-u}(\xi_u - \sigma T S_u)\mathrm{d}u. \tag{4.36}$$

After further substitution and rearrangement of Eq. (4.36) one can show that

$$\tilde{B}_s - \tilde{B}_t = B_s - B_t + \int_t^s \frac{\sigma T}{T-u}(X_T - S_u)\mathrm{d}u\ , \tag{4.37}$$

implying that the increment $B_s - B_t$ is independent of the following σ−algebra $\left\{\mathcal{F}_t^{\xi}\right\} = \sigma\left(\{\xi_t\}_{0\leq s<t}\right)$ and that

$$\mathbb{E}^{\mathbb{Q}}\left[\tilde{B}_s - \tilde{B}_t \middle| \mathcal{F}_t^{\xi}\right] = \mathbb{E}^{\mathbb{Q}}\left[\tilde{B}_s - \tilde{B}_t \middle| \xi_t\right] = 0 \tag{4.38}$$

as it can be shown that

$$\mathbb{E}^{\mathbb{Q}}\left[\int_t^s \frac{\sigma T}{T-u}(X_T - S_u)\mathrm{d}u \middle| \xi_t\right] = 0. \tag{4.39}$$

So to conclude the proof, Eq. (4.38) shows that the two standard Brownian motion in question are independent, that is $\tilde{B}_t \perp\!\!\!\perp^{\mathbb{Q}} B_t$.

Second to prove (ii): if $\{\tilde{B}_t\}$ in Eq. (4.33) is an martingale then

$$\mathbb{E}\left[\tilde{B}_\tau \mid \mathcal{F}_t^\xi\right] = \tilde{B}_t \quad \text{for } 0 \le t \le \tau < T \tag{4.40}$$

and hence is an $\{\mathcal{F}_t^\xi\}$-Brownian motion. Integrating Eq. (4.33)

$$\tilde{B}_t = \int_0^t \frac{1}{T-u}(\xi_u - \sigma T S_u)\mathrm{d}u + \xi_t \tag{4.41}$$

$$\begin{aligned}\mathbb{E}\left[\tilde{B}_\tau - \tilde{B}_t \mid \mathcal{F}_t^\xi\right] = \mathbb{E}\Bigg[& \int_0^\tau \tfrac{1}{T-u}(\xi_u - \sigma T S_u)\mathrm{d}u + \xi_\tau \\ & - \int_0^t \tfrac{1}{T-u}(\xi_u - \sigma T S_u)\mathrm{d}u + \xi_t \Bigg| \mathcal{F}_t^\xi \Bigg]\end{aligned} \tag{4.42}$$

$$\begin{aligned}\mathbb{E}\left[\tilde{B}_\tau \mid \mathcal{F}_t^\xi\right] = \tilde{B}_t - \mathbb{E}\left[\xi_t \mid \mathcal{F}_t^\xi\right] + \mathbb{E}\left[\xi_\tau \mid \mathcal{F}_t^\xi\right] - \sigma T \mathbb{E}\left[\int_t^\tau \tfrac{1}{T-u} S_u \mathrm{d}u \middle| \mathcal{F}_t^\xi\right] \\ + \mathbb{E}\left[\int_t^\tau \tfrac{1}{T-u} \xi_u \mathrm{d}u \middle| \mathcal{F}_t^\xi\right].\end{aligned} \tag{4.43}$$

Since the information process $\{\xi_t\}_{t\ge 0}$ is a Markov process shown in Eq. (4.18), then the expectation in Eq. (4.43) becomes $\mathbb{E}\left[. \mid \mathcal{F}_t^\xi\right] = \mathbb{E}\left[. \mid \xi_t\right]$, so

$$\begin{aligned}\mathbb{E}\left[\tilde{B}_\tau \mid \xi_t\right] = \tilde{B}_t - \mathbb{E}\left[\xi_t \mid \xi_t\right] + \mathbb{E}\left[\xi_\tau \mid \xi_t\right] - \sigma T \mathbb{E}\left[\int_t^\tau \tfrac{1}{T-u} \mathbb{E}[X_T \mid \xi_u] \mathrm{d}u \middle| \xi_t\right] \\ + \mathbb{E}\left[\int_t^\tau \tfrac{1}{T-u} \left(\sigma X_T u + \beta_{uT}\right) \mathrm{d}u \middle| \xi_t\right].\end{aligned} \tag{4.44}$$

Applying the tower property and integrating out the terms in Eq. (4.44) one finds

$$\begin{aligned}\mathbb{E}\left[\tilde{B}_\tau \mid \xi_t\right] = {} & \tilde{B}_t - \mathbb{E}\left[\sigma X_T t \mid \xi_t\right] - \mathbb{E}\left[\beta_{tT} \mid \xi_t\right] + \mathbb{E}\left[\sigma X_T \tau + \beta_{\tau T} \mid \xi_t\right] \\ & - \sigma T \mathbb{E}[X_T \mid \xi_t] \left(-\ln\left(\left|\frac{\tau - T}{t - T}\right|\right)\right) + \sigma \mathbb{E}\left[X_T \mid \xi_t\right] \left(-T \ln\left(\left|\frac{\tau - T}{t - T}\right|\right) - \tau + t\right) \\ & + \int_t^\tau \tfrac{1}{T-u} \mathbb{E}\left[\beta_{uT} \mid \xi_t\right] \mathrm{d}u\,;\end{aligned} \tag{4.45}$$

using the tower property again for the term $\mathbb{E}\left[\beta_{\tau T} \mid \xi_t\right] = \mathbb{E}\left[\mathbb{E}\left[\beta_{\tau T} \mid \beta_{tT}\right] \mid \xi_t\right]$, one can cancel terms such that

$$\mathbb{E}\left[\tilde{B}_\tau \mid \xi_t\right] = \tilde{B}_t - \mathbb{E}[\underbrace{\mathbb{E}\left[\beta_{tT} \mid \beta_{tT}\right]}_{=0} \mid \xi_t] + \mathbb{E}\left[\mathbb{E}\left[\beta_{\tau T} \mid \beta_{tT}\right] \mid \xi_t\right] \\ + \int_t^\tau \tfrac{1}{T-u}\mathbb{E}\left[\mathbb{E}\left[\beta_{uT} \mid \beta_{tT}\right] \mid \xi_t\right] \mathrm{d}u \, . \tag{4.46}$$

The final step is to apply the following relationship $\mathbb{E}\left[\beta_{\tau T} \mid \xi_t\right] = \frac{T-\tau}{T-t}\beta_{tT}$ leading to

$$\mathbb{E}\left[\tilde{B}_\tau \mid \xi_t\right] = \tilde{B}_t \underbrace{- \frac{T-\tau}{T-t}\beta_{tT} + \frac{\tau - T}{T-t}\beta_{tT}}_{=0} \tag{4.47}$$

which is the desired result that $\mathbb{E}\left[\tilde{B}_\tau \mid \xi_t\right] = \tilde{B}_t$. Therefore we have proved that the process $\left\{\tilde{B}_t\right\}$ is a martingale with respect to the information generated by $\{\xi_t\}$ and is an $\left\{\mathcal{F}_t^\xi\right\}$-Brownian motion.

What the BHM model shows from Eq. (4.32) is that the standard Brownian $\left\{\tilde{B}_t\right\}$ drives the price process with noise. The noise is thus driven by information through t

$$\zeta_{tT} = \frac{\sigma T}{T-t}V_t \tag{4.48}$$

where the price volatility is a function of time and $\{\xi_t\}$, denoted as $\zeta_t \equiv \Sigma(\xi_t, t)$. Combining Eq. (4.32) and Eq. (4.48) one finds that the price forms the drift free Itô process

$$\mathrm{d}S_t = \zeta_t \mathrm{d}\tilde{B}_t \tag{4.49}$$

and if $\{S_t\}$ is bounded that is $\mathbb{E}^{\mathbb{Q}}\left[\left(\int_0^T \zeta_t^2 \mathrm{d}t\right)^{1/2}\right] < \infty$ then it is a martingale with respect to the risk-neutral measure $\mathbb{Q}$. This Eq. (4.49) is an important break though because as noted by Brody-Hughston-Macrina in [86]

> *"In this way our framework resolves the paradoxical point of view usually adopted in financial modelling in which $\{W_t\}$ is regarded on the one hand as "noise", and yet on the other hand also generates the market information flow."*

which means price can now be viewed as an emergent phenomenon that is driven by the noisiness of the flow of information. Consider the example: if σ is large then the market more or less knows the price and the noise is small, and compare that to the situation when σ is small: the price process is driven by pure noise and the market cannot be sure of the price.

4.2.3 Binary State X_T

It is known that if the random variable X_T has a discrete outcome with two values, such as the binary set, the price Eq. (4.25) has a closed form solution. This section shows that if the final state is defined as $X_T \in \{0, 1\}$ the price process can be analytically solved. Setting the sample space of the random variable such that $\{\omega \in \Omega \mid \Omega = \{\text{Payment}, \text{Non-Payment}\}\}$ which has the following map

$$X_T(\omega) = \begin{cases} 1, & \text{if } \omega = \text{Payment}, \\ 0, & \text{if } \omega = \text{Non-Payment}, \end{cases} \tag{4.50}$$

and the a priori probabilities

$$f(x) = \begin{cases} p_1, & \text{if } x = x_1 = 1, \\ p_0, & \text{if } x = x_0 = 0, \end{cases} \tag{4.51}$$

such that $p_0 + p_1 = 1$ and $p_0, p_1 \in (0, 1)$. The conditional pricing density which was found in Eq. (4.24) can be adjusted for a discrete x or in this case a binary one, giving

$$\phi_t(\{x_i\}_{i=0,1}) = \frac{p_i \exp\left(\frac{T}{T-t}\sigma x_i(\xi_t - \frac{1}{2}\sigma x_i t)\right)}{\sum_{j=0}^{1} p_j \exp\left(\frac{T}{T-t}\sigma x_j(\xi_t - \frac{1}{2}\sigma x_j t)\right)} \tag{4.52}$$

such that the integral in the numerator has now changed to a sum $\left(\int_0^1 \to \sum_0^1\right)$. Denoting the conditional density function as $\phi_t(\{x_i\}_{i=0,1}) \triangleq \phi_{it}$, to stream line the notation, one finds that Eq. (4.52) becomes

$$\phi_{it} = \frac{p_i \exp\left(\frac{T}{T-t}\sigma x_i(\xi_t - \frac{1}{2}\sigma x_i t)\right)}{\left(p_0 + p_1 \exp\left(\frac{T}{T-t}\sigma(\xi_t - \frac{1}{2}\sigma t)\right)\right)}\,, \tag{4.53}$$

where $\phi_{0t} + \phi_{1t} = 1$. The conditional density function $\{\phi_{it}\}$ is a stochastic process, the dynamics of which drive the dynamics of the price process. Using the result found in Eq. (4.53) one can differentiate the process $\{\phi_{it}\}$ and find

$$\frac{\mathrm{d}\phi_{it}}{\phi_{it}} = \frac{\sigma T}{T-t}\left(x_i - \phi_{1t}\right)\left(\frac{1}{T-t}(\xi_t - \sigma T\phi_{1t})\mathrm{d}t + \mathrm{d}\xi_t\right), \tag{4.54}$$

where $\phi_{1t} = S_t$, which is shown below in Eq. (4.57). Combining the result Eq. (4.54) with Eq. (4.33) the dynamic process of $\{\phi_{it}\}$ satisfies the following diffusion equation

$$\frac{\mathrm{d}\phi_{it}}{\phi_{it}} = \frac{\sigma T}{T-t}\left(x_i - \phi_{1t}\right)\mathrm{d}\tilde{B}_t \tag{4.55}$$

where $\left\{\tilde{B}_t\right\}$ is a standard Brownian motion. The price is found by using the discrete conditional expectation Eq. (4.25)

$$S_t = \sum_{i=0}^{1} x_i\phi_{it} = \left(1 + \tfrac{p_0}{p_1}\exp\left(-\tfrac{T\sigma}{T-t}(\xi_t - \tfrac{1}{2}\sigma t)\right)\right)^{-1} \tag{4.56}$$

where $\frac{p_0}{p_1}$ is the ratio of the a priori probabilities, that is $\mathbb{Q}\left[X_T = 0\right]/\mathbb{Q}\left[X_T = 1\right]$ for $t = 0$. The compact solution Eq. (4.56) shows that the price of the binary X_T is conveniently

$$S_t = \mathbb{E}^{\mathbb{Q}}[X_T|\xi_t] = \phi_{1t} \tag{4.57}$$

which is plotted in the top plot of Fig. 4.2. Combining the result Eq. (4.55) with the price Eq. (4.57) one finds the following dynamic price process

$$\mathrm{d}S_t = \sum_{i=0}^{1} x_i\mathrm{d}\phi_{it} = \frac{\sigma T}{T-t}V_t\left(\frac{1}{T-t}(\xi_t - \sigma T\phi_{1t})\mathrm{d}t + \mathrm{d}\xi_t\right), \tag{4.58}$$

which is the same result as found in Eq. (4.30), where V_t is the conditional variance process $\{V_t\}$ and is defined to be

$$V_t = (1 - \phi_{1t})\phi_{1t}\ , \tag{4.59}$$

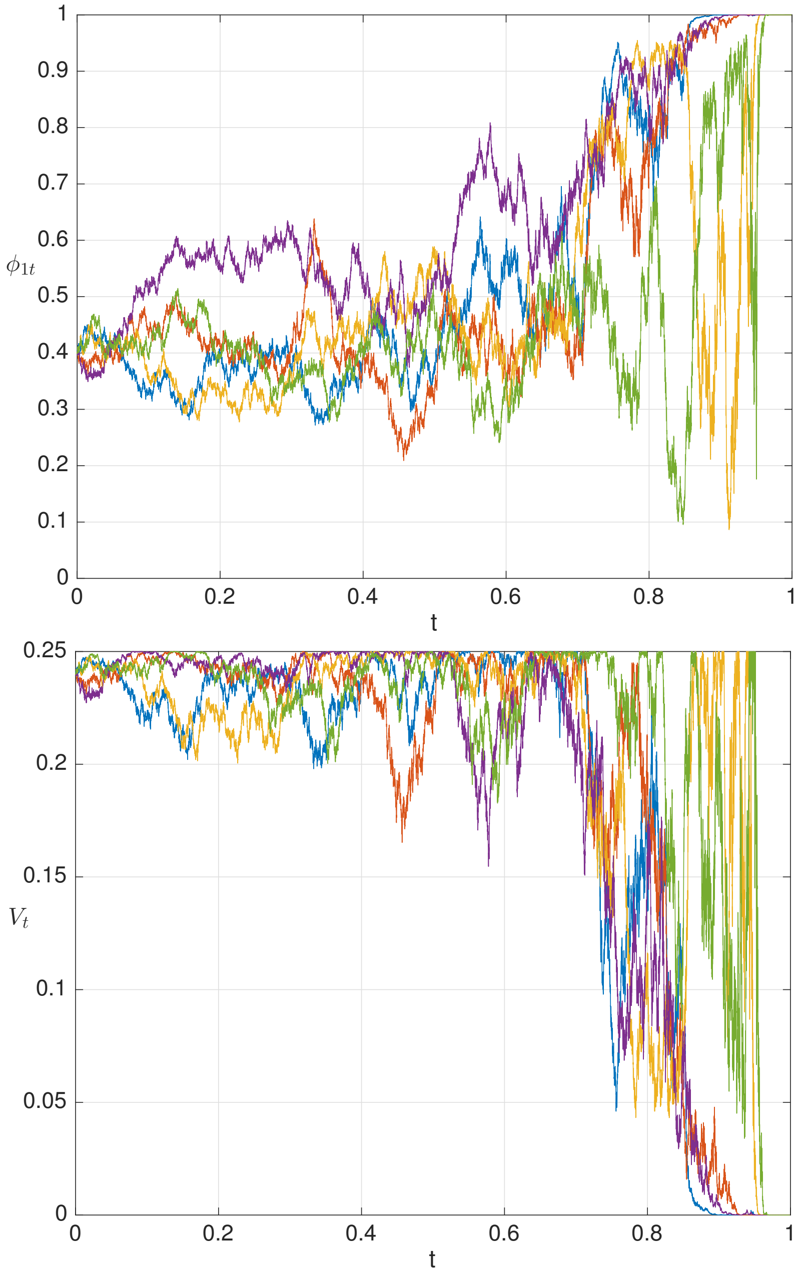


Figure 4.2: Five independent price processes as defined in Eq. (4.57) in the upper plot with corresponding conditional variance as defined in Eq. (4.59) in the lower plot with the parameters; time step $\Delta t = 1 \times 10^{-5}$, $T = 1$, $\sigma = 1$, $p_1 = 0.4$ and $X_T = 1$.

which is plotted in the bottom plot of Fig. 4.2, and is quadratic in ϕ_{1t} with roots at 0 and 1. Using Eq. (4.33) one can write the dynamic price process as

$$\mathrm{d}S_t = \frac{\sigma T}{T-t}(1-\phi_{1t})\phi_{1t}\mathrm{d}\tilde{B}_t \tag{4.60}$$

where $\{\tilde{B}_t\}$ is standard Brownian motion. The same definition of volatility as in Eq. (4.48), but for binary states, gives

$$\zeta_{tT} \triangleq \frac{\sigma T}{T-t}(1-\phi_{1t})\phi_{1t} = \frac{\sigma T}{T-t}V_t, \tag{4.61}$$

and this leads to a final result

$$\mathrm{d}S_t = \zeta_{tT}\mathrm{d}\tilde{B}_t. \tag{4.62}$$

which is a diffusion equation similar to that found in Eq. (4.49). The results outlined in this section are fundamental in the development of our trading model is Sec. 4.3 and to the calibration of this model to real data in Sec. 4.7.

4.2.4 Shannon Entropy

To investigate the evolution and dynamics of the uncertainty in the price process $\{S_t\}$, one can evaluate the statistical measure known as Shannon entropy [127, 128]. The Shannon entropy represents a measure of the expected informational content of a random variable that has been lost, and there is an uncertainty about the value of the random variable. As uncertainty decreases in the random variable the Shannon entropy must be decreasing, which makes it similar to a statistical measure such as variance.

Consider a $n+1$ discrete cash flow state denoted as $X_T \in \{x_0, x_1, \ldots, x_n\}$ which has the same probability density function as defined in Eq. (4.53), but instead of a binary set we now take $i = 0, 1, \ldots, n$ states for the process $\{\phi_{it}\}$. The time dependent Shannon entropy for this process $\{\phi_{it}\}$ is

$$H_t \triangleq \mathbb{E}^{\phi_{it}}\left[-\ln\left(\phi_{it}\right)\right] = -\sum_{i=0}^{n} \phi_{it} \ln\left(\phi_{it}\right) \ , \tag{4.63}$$

thus for the binary system $X_T \in \{x_0, x_1\}$ the Shannon entropy is

$$H_t = -\sum_{i=0}^{1} \phi_{it} \ln(\phi_{it}) = -\phi_{0t} \ln(\phi_{0t}) - \phi_{1t} \ln(\phi_{1t}) \ . \tag{4.64}$$

where $H_T = 0$; as one is certain of the value of X_T at $t = T$. The Shannon entropy H_t has been plotted in Fig. 4.3. The initial value of Shannon entropy H_{0T} can be found from the fact that $\phi_{i0} = p_i$ for $i = 0, 1$, therefore $H_{0T} = -\left(p_0 \ln(p_0) + p_1 \ln(p_1)\right)$. Differentiating the time dependent entropy Eq. (4.64) and setting $X_T \in \{0, 1\}$, one finds the dynamic process

$$\mathrm{d}H_t = -\frac{\sigma T}{T-t}\phi_{1t}\left(\frac{\sigma T}{2(T-t)}(1-\phi_{1t})\mathrm{d}t + (\ln\phi_{1t} + H_t)\,\mathrm{d}\tilde{B}_t\right) \tag{4.65}$$

which has a drift that is strictly negative which arises from the fact that the conditional variance V_t is a supermartingale. Remembering that the variance

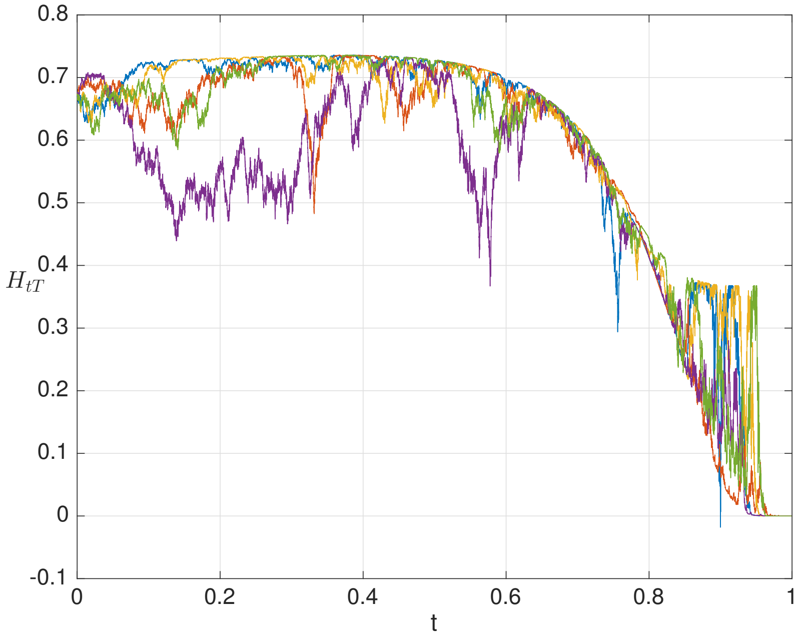


Figure 4.3: Five independent Shannon entropy processes as defined in Eq. (4.64) with the parameters; time step $\Delta t = 1 \times 10^{-5}$, $T = 1$, $\sigma = 1$, $p_1 = 0.4$ and $X_T = 1$.

process V_t is given by Eq. (4.59) and integrating Eq. (4.65)

$$H_t - H_0 = -\int_0^t \frac{\sigma^2 T^2}{2(T-u)^2} V_u \mathrm{d}u - \int_0^t h(\phi_{1u}, u) \mathrm{d}\tilde{B}_u \tag{4.66}$$

where we have defined $h(\phi_{1t}, t) = \frac{\sigma T}{T-t} \phi_{1t} \left(\ln \phi_{1t} + H_t\right)$. Taking the expectation with respect to $\mathbb{Q}$, the risk-neutral measure, we find that $\mathbb{E}^{\mathbb{Q}}\left[\int_0^t h(\phi_{1u}, u) \mathrm{d}\tilde{B}_u\right] = 0$ for $\forall t \in [0, T]$ because of the martingale property (not proven here but see [90] for proof). Thus, we are left with following expression for the expected Shannon entropy

$$\mathbb{E}^{\mathbb{Q}}\left[H_t\right] = H_0 - \mathbb{E}^{\mathbb{Q}}\left[\int_0^t \frac{\sigma^2 T^2}{2(T-u)^2} V_u \mathrm{d}u\right] \tag{4.67}$$

where the expectation term on the right is known as the *mutual information* between the information process ξ_t and the terminal payoff X_T [90, 128]. Taking the time limit $t \to T$ and using the fact that the entropy process converges to zero as $t \to T$ then Eq. (4.67) converges in the same way, hence

$$\lim_{t\to T} \mathbb{E}^{\mathbb{Q}}\left[H_0 - \int_0^t \frac{\sigma^2 T^2}{2(T-u)^2} V_u \,\mathrm{d}u\right] = \lim_{t\to T} \mathbb{E}^{\mathbb{Q}}\left[H_0 - \int_0^t \frac{\sigma T}{2(T-u)} \zeta_{uT} \,\mathrm{d}u\right] = 0 \,, \tag{4.68}$$

where ζ_{tT} is the volatility process defined in Eq. (4.61). This relation shows that for the binary $X_T \in \{0,1\}$ the integration constant H_0 is equal to the following relationship

$$H_0 = \mathbb{E}^{\mathbb{Q}}\left[\int_0^T \frac{\sigma T}{2(T-u)^2} V_{uT} \,\mathrm{d}u\right] = \mathbb{E}^{\mathbb{Q}}\left[\int_0^T \frac{\sigma T}{2(T-u)} \zeta_{uT} \,\mathrm{d}u\right] , \tag{4.69}$$

which is bound by the value $\ln(2)$. Therefore, Eq. (4.69) has the following interpretation that: the variance V_t and volatility ζ_{tT} process must decay in time sufficiently quickly to ensure the right hand side of Eq. (4.69) is to exist. Furthermore, the results Eq. (4.68) and Eq. (4.69) highlight that the price dynamics of the binary bond, as found in Eq. (4.62), are such that the volatility ζ_{tT} of S_t converges as $\lim_{t\to T} \zeta_{tT} = 0$, which ensures that any divergence created by the $(T-t)^{-1}$ term in Eq. (4.61) is rebalanced as H_0 is bounded.

The combination of all these entropic properties found in this section lead to an important property of the conditional density function $\{\phi_{it}\}$, Eq. (4.53), that converges as

$$\lim_{t\to T} \phi_{it} = \mathbb{1}_{\{X_T = i\}} \text{ for } i = 0, 1 \,, \tag{4.70}$$

where $\mathbb{1}$ is the indicator function. This final result Eq. (4.70), ensures that price, as calculated in Eq. (4.60), converges to the desired terminal value $\lim_{t\to T} S_t = X_T$.

4.3 Trading

This section is the basis for the following working paper [129] which explores a numerical model for an agent based market where the agents perceive prices with their own BHM pricing kernel. Trading is an important mechanism within financial markets, and without it, market prices could not form. What this section will try and achieve is to take the binary state $X_T \in \{0,1\}$ pricing model derived in

Sec. 4.2.3 and introduce a mechanism where agents can buy or sell such an asset. A model was briefly introduced in the article [94] but was never simulated; this section moves in a similar spirit by creating trading agents but not exactly in the same way. The difference in our approach is that we create a mechanism where the trading agents have a price shift and this update their information process, where as in [94] they update the price directly by updating the information process only.

Given that the price process calculated in Eq. (4.60) is driven by the noise process ξ_t one can parametrise a model that consists of two agents who are pricing the same asset. The two agents have their own perceptions of the asset X_T's price which is modelled with the following two information processes

$$\begin{aligned} \xi_t^{(1)} &= \sigma_1 t X_T + \beta_{tT}^{(1)} \\ \xi_t^{(2)} &= \sigma_2 t X_T + \beta_{tT}^{(2)} \end{aligned} \tag{4.71}$$

where the two information rate constants are defined to be strictly positive $0 < \sigma_1 < \infty$, $0 < \sigma_2 < \infty$. The time domain is $t \in [0, T]$ and the standard Brownian bridge processes are independent under the risk-neutral measure $\mathbb{Q}$, denoted as $\beta_{tT}^{(1)} \perp\!\!\!\perp^{\mathbb{Q}} \beta_{tT}^{(2)}$. Applying the rules as set out in Sec. 4.2.3, the pricing kernel is Eq. (4.19) which is assumed to give the two mid-prices for each of the two agents namely

$$\begin{aligned} S_t^{(1)} &= \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t^{(1)}\right] = \phi_{1t}^{(1)} \\ S_t^{(2)} &= \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t^{(2)}\right] = \phi_{1t}^{(2)} \end{aligned} \tag{4.72}$$

which are driven endogenously by their individual information processes Eq. (4.71). The most competitive sell and buy prices are defined in such away that

$$S_{A/B}^{(i)} = S_t^{(i)} \pm \delta \tag{4.73}$$

where the agent index is $i = 1, 2$, the subscript A is the *ask* price (*sell* price) and B is the *bid* price (*buy* price). For simplicity, the model will consider a constant*

*δ being a positive constant is only assumed for simplicity but it can be represented as a convex function $\delta(I)$ where I is measure of inventory for a particular agent.

half spread δ, where this parameter controls the rate of trading between trader 1 and 2.

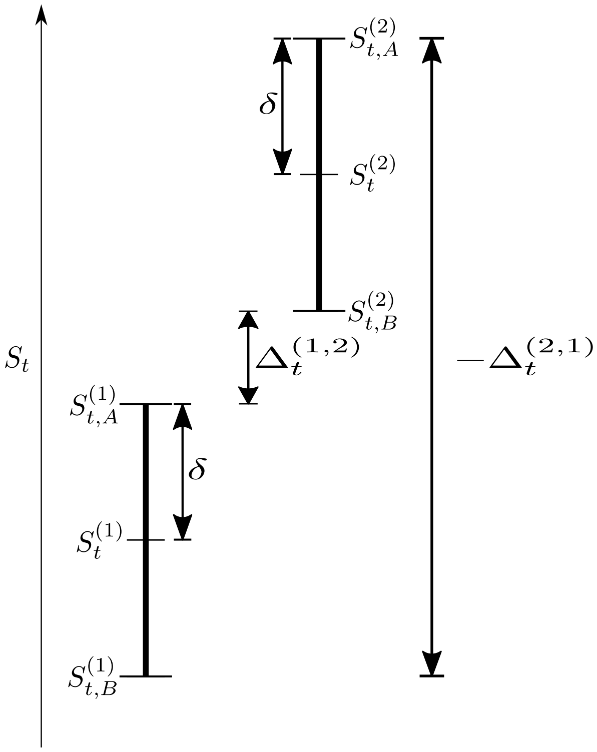


Figure 4.4: This diagram outlines the geometry in trading mechanism used throughout the model and relationship between the parameters: half spread δ, mid-price $S_t^{(\cdot)}$, ask price $S_{t,A}^{(\cdot)}$, bid price $S_{t,B}^{(\cdot)}$, and the disagreement in bid and ask price $\Delta_t^{(\cdot,\cdot)}$.

The difference between the two agents' *bid* and *ask* price is defined as

$$\Delta_t^{(i,j)} \triangleq S_{t,B}^{(i)} - S_{t,A}^{(j)} \tag{4.74}$$

where $i, j = 1, 2$ and $i \neq j$: when $i = j$ the spread is simply $\Delta_t^{(i,i)} = \Delta_t^{(j,j)} = -2\delta$.

Hence the spreads can be represented in the following matrix[†] with

$$\bar{\bar{\Lambda}} = \begin{pmatrix} -2\delta & \Delta_t^{(1,2)} \\ \Delta_t^{(2,1)} & -2\delta \end{pmatrix} \tag{4.75}$$

where the determinant is defined as $\det\left(\bar{\bar{\Lambda}}\right) = 4\delta^2 - \Delta_t^{(2,1)}\Delta_t^{(1,2)}$. It can seen from Fig. (4.4) that when $\Delta_t^{(1,2)} = 0$, $\Delta_t^{(2,1)} = -4\delta$ or $\Delta_t^{(2,1)} = 0$, $\Delta_t^{(1,2)} = -4\delta$ and when $\Delta_t^{(1,2)} \neq 0$ or $\Delta_t^{(2,1)} \neq 0$ then $\Delta_t^{(2,1)}\Delta_t^{(1,2)} < 0$, which means that

$$\det\left(\bar{\bar{\Lambda}}\right) = \begin{cases} 4\delta^2, & \text{if } S_{t,B}^{(i)} = S_{t,A}^{(j)} \quad \forall i,j \\ \left(2\delta + \Delta_t^{(i,j)}\right)^2, & \text{if } S_{t,B}^{(i)} < S_{t,A}^{(j)} \text{ where } i \neq j \end{cases} \tag{4.76}$$

which shows that $\det\left(\bar{\bar{\Lambda}}\right) \geq 4\delta^2 > 0$ if and only if $\delta > 0$, and $\det\left(\bar{\bar{\Lambda}}\right) \geq 0$ if and only if $\delta \geq 0$. The eigenvalues of $\bar{\bar{\Lambda}}$ if $\Delta > 0$ are found to be $\lambda = -2\delta \pm \sqrt{4\delta^2 - \det\left(\bar{\bar{\Lambda}}\right)}$ $\Rightarrow \lambda = -2\delta \pm \sqrt{\Delta_t^{(1,2)}\Delta_t^{(2,1)}}$ with eigenvectors $z = 1 \pm \sqrt{\frac{\Delta_t^{(1,2)}}{\Delta_t^{(2,1)}}}$ giving possible two outcomes:

$$\begin{aligned} &(i)\,\mathrm{Im}(\lambda) = \mathrm{Im}(z) = 0 && \text{if } \Delta_t^{(1,2)} \leq 0 \text{ or } \Delta_t^{(2,1)} \leq 0 \\ &(ii)\,\mathrm{Im}(\lambda) \neq 0 \textit{ and } \mathrm{Im}(z) \neq 0 && \text{if } \Delta_t^{(1,2)} > 0 \text{ or } \Delta_t^{(2,1)} > 0\,. \end{aligned} \tag{4.77}$$

This is the first check to see if a trade is possible where outcome (i) means it is feasible or outcome (ii) is not feasible at all. This setup gives a trigger mechanism for when a trade can be initiated, and put simply a trade is possible if the two ranges of price quoted by the two agents do not overlap. If the two agents *disagree* in their pricing then a trade is possible if

$$\Delta_t^{(i,j)} > 0 \Rightarrow S_A^{(i)} > S_B^{(j)} \tag{4.78}$$

where $i, j = 1, 2$ and $i \neq j$ and the time of the trade is recorded and denoted as t^*. The next step of the trade would be for the two participants in the trade to

[†] The matrix $\bar{\bar{\Lambda}}$ Eq. (4.75) is a special case of a non-symmetric real matrix that is negative-definite if $\delta > 0$ and negative-semidefinite if $\delta \geq 0$.

update their prices by the size of the disagreement which is defined as

$$\Theta_t^{(i,j)} \hat{=} \delta + \frac{\left|\Delta_t^{(i,j)}\right|}{2} \tag{4.79}$$

which is shown in Fig. 4.5. The variable $\Theta_t^{(i,j)}$ is equivalent to the distance each agent needs to move to adjust each their price to the average of the *bid-ask* price. The direction of the change in price adjustment for the agent depends on whether the agent is buying or selling and can be described by

$$\begin{aligned} \Delta S_t^{(1)} &= -\Theta_t^{(2,1)} H\left(\Delta^{(2,1)}\right) + \Theta_t^{(1,2)} H\left(\Delta^{(1,2)}\right) \ , \\ \Delta S_t^{(2)} &= \quad \Theta_t^{(2,1)} H\left(\Delta^{(2,1)}\right) - \Theta_t^{(1,2)} H\left(\Delta^{(1,2)}\right), \end{aligned} \tag{4.80}$$

where $H(.)$ is the Heaviside step function defined as

$$H(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x \le 0, \end{cases} \tag{4.81}$$

where $x \in \mathbb{R}$. The change in price in Eq. (4.80) has the following structure when a trade is feasible: $\Delta S_{t^*}^{(1)} < 0 \Leftrightarrow \Delta S_{t^*}^{(2)} > 0$ where the absolute values $\left|\Delta S_{t^*}^{(1)}\right| = \left|\Delta S_{t^*}^{(2)}\right| = \Theta_{t^*}^{(1,2)}$ or $\Delta S_{t^*}^{(1)} > 0 \Leftrightarrow \Delta S_{t^*}^{(2)} < 0$ where $\left|\Delta S_{t^*}^{(1)}\right| = \left|\Delta S_{t^*}^{(2)}\right| = \Theta_{t^*}^{(2,1)}$ otherwise $\Delta S_t^{(1)} = \Delta S_t^{(2)} = 0$. This gives the correct structure to the agents' price adjustment: when agents buy the asset there is update in the up direction and if the agent sells there is a shift in the down direction. The corresponding price shifts $\Delta S_{t^*}^{(1)}$ and $\Delta S_{t^*}^{(2)}$ can be used to calculate the changes in information $\Delta\xi_{t^*}^{(1)}$ and $\Delta\xi_{t^*}^{(2)}$ which would be needed to bring about these changes $\Delta S_{t^*}^{(1)}$ and $\Delta S_{t^*}^{(2)}$. These changes are found by inverting Eq. (4.56), hence the changes in information $\Delta\xi_{t^*}^{(1)}$ and $\Delta\xi_{t^*}^{(2)}$ needed to update the prices by $\Delta S_{t^*}^{(1)}$ and $\Delta S_{t^*}^{(2)}$ are respectively

$$\begin{aligned} \Delta\xi_{t^*}^{(1)} &= \frac{1}{2}\sigma_1 t^* - \xi_{t^*}^{(1)} - \left(\frac{T - t^*}{T\sigma_1}\right)\ln\left(\frac{p_1}{p_0}\left(\frac{1}{S_{t^*}^{(1)} + \Delta S_{t^*}^{(1)}}\right)\right), \\ \Delta\xi_{t^*}^{(2)} &= \frac{1}{2}\sigma_2 t^* - \xi_{t^*}^{(2)} - \left(\frac{T - t^*}{T\sigma_2}\right)\ln\left(\frac{p_1}{p_0}\left(\frac{1}{S_{t^*}^{(2)} + \Delta S_{t^*}^{(2)}}\right)\right), \end{aligned} \tag{4.82}$$

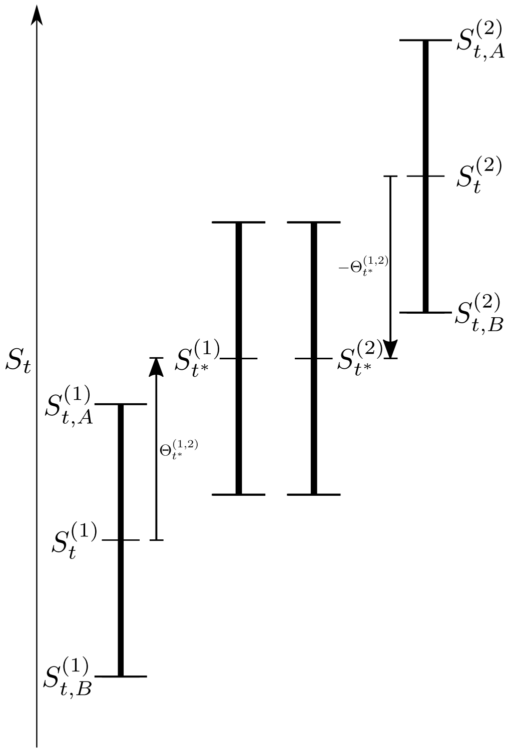


Figure 4.5: The notation in this diagram is same as Fig. 4.4 but here we show the mechanism of the price updating after a trade is trigged. The mid-prices $S^{(1)}_t$ and $S^{(2)}_t$ are shifted to the updated to prices $S^{(1)}_{t^*}$ and $S^{(2)}_{t^*}$ respectively. Agent 1 is the selling agent and agent 2 is the buying agent. The quantity $\Theta^{(1,2)}_{t^*}$ is the amount both agents have to adjust to move to their average price and the arrow is the direction in which they adjust; positive for up shift (sell) and negative for down shift (buy).

which is effectively the information obtained by each agent through having made a trade. We can now redefine the two information processes Eq. (4.71) to account for trading jumps

$$\begin{aligned} \xi'^{(1)}_t &= \left(\sigma_1 X_T - \sum_{t^* \in \mathbb{T}^*} \frac{1}{T - t^*} \Delta\xi^{(1)}_{t^*} H(t^*) \right) t + \beta^{(1)}_{tT} \\ \xi'^{(2)}_t &= \left(\sigma_2 X_T - \sum_{t^* \in \mathbb{T}^*} \frac{1}{T - t^*} \Delta\xi^{(2)}_{t^*} H(t^*) \right) t + \beta^{(2)}_{tT} \end{aligned} \tag{4.83}$$

where $H(.)$ the Heaviside function is used to ensure that trading updates only happen when $t > t^*$. The term $\sum_{t^* \in \mathbb{T}^*} \frac{1}{T-t^*} \Delta\xi_{t^*}^{(i)}$ adds the necessary jumps guaranteeing that the information process, future trajectory is corrected linearly ensuring that the prices remain in the unit square $S_t^{(i)} \in [0,1] \times [0,1]$ for $i = 1, 2$. Thus, the pricing of the asset is now calculated as

$$\begin{aligned} S_t^{(1)} &= \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t'^{(1)}\right] = \phi_{1t}^{(1)} \\ S_t^{(2)} &= \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t'^{(2)}\right] = \phi_{1t}^{(2)} \end{aligned} \tag{4.84}$$

and the dynamics are found in the same way as in Eq. (4.60) but with the updated information processes $\xi_t'^{(1)}$ and $\xi_t'^{(2)}$.

4.3.1 Trading with N-agents

In the previous section, Sec. 4.3, only two agents were used but the model can easily be expanded to any number of agents which is denoted as N. This means that there are N information processes implying that Eq. (4.71) and Eq. (4.72) become the following set of process

$$\begin{matrix} \xi_t^{(1)} = \sigma_1 t X_T + \beta_{tT}^{(1)} & & S_t^{(1)} = \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t^{(1)}\right] = \phi_{1t}^{(1)} \\ \xi_t^{(2)} = \sigma_2 t X_T + \beta_{tT}^{(2)} & & S_t^{(2)} = \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t^{(2)}\right] = \phi_{1t}^{(2)} \\ \vdots & \Rightarrow & \vdots \\ \xi_t^{(N)} = \sigma_N t X_T + \beta_{tT}^{(N)} & & S_t^{(N)} = \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t^{(N)}\right] = \phi_{1t}^{(N)} \end{matrix} \tag{4.85}$$

where the information rate constants are defined to be strictly positive $0 < \sigma_1 < \infty$, $0 < \sigma_2 < \infty$,...,$0 < \sigma_N < \infty$. The time domain is $t \in [0, T]$ and the standard Brownian bridge processes are all independent of each other. The definition of the *bid* and *ask* price remains the same as in Eq. (4.73) and the distance between different agent's buy and sell price is Eq. (4.74). The matrix $\bar{\bar{\Lambda}}$, defined in Eq. (4.75), now becomes a block matrix of 2×2 matrices. This block matrix can defined as the following upper triangular matrix

$$\Xi = \begin{pmatrix} \bar{\bar{\Lambda}}_{1,1} & \bar{\bar{\Lambda}}_{1,2} & \bar{\bar{\Lambda}}_{1,3} & \dots & \bar{\bar{\Lambda}}_{1,N} \\ & \bar{\bar{\Lambda}}_{2,2} & \bar{\bar{\Lambda}}_{2,3} & \dots & \bar{\bar{\Lambda}}_{2,N} \\ & & \bar{\bar{\Lambda}}_{3,3} & \dots & \bar{\bar{\Lambda}}_{3,N} \\ & & & \ddots & \vdots \\ & & & & \bar{\bar{\Lambda}}_{N,N} \end{pmatrix} \tag{4.86}$$

where the elements (which are 2 × 2 matrices) are defined as the following

$$\bar{\bar{\Lambda}}_{i,j} = \begin{pmatrix} -2\delta & \Delta_t^{(i,j)} \\ \Delta_t^{(j,i)} & -2\delta \end{pmatrix} \tag{4.87}$$

and $\Delta_t^{(i,j)}$ are defined as in Eq. (4.74). The diagonal elements of Eq. (4.86), that is $i = j$, are of no interest to trading as agents can not trade with themselves. The upper off-diagonal elements determine whether a trade is feasible between the two agents $i \neq j$: there are $N(N-1)/2$ matrices that need to be checked if the conditions from Eq. (4.77) are feasible, hence the time overhead of the number of checks grows as $\mathcal{O}(N^2)$.

4.3.2 Market Structure

This section looks at how the model creates a market with three different types of traders. The market configurations are denoted as (N_{MM}, N_{IT}, N_{NT}) where N_{MM} is the number of market makers, N_{IT} is the number of informed traders and N_{NT} is the number of noise traders. The different traders in the model are summarised as following:

1. Market makers are traders that define the spread by quoting the buy and sell prices.

2. Informed traders do not agree with the market and believe in their own price.

3. Noise traders are agents only looking to make a trade for exogenous reasons.

These three different types of traders will be described in more detail in the next few sections.

4.3.3 Market Maker

A market maker has the role (or strategy) in financial markets to provide liquidity to exchanges. This strategy has an inherent risk associated with it, as the market maker bears the risk of holding sometimes large quantities of financial assets. The risk of holding financial assets emerges because their prices can adversely move against the holder; for example the price of a stock, that is held, could move down. A market maker looks to make the bid-ask spread of an asset by holding in the hope to buy the asset for a lower price at which they can broker a deal to sell it. In our model of trading we have thus defined a trading trigger for when this type of trade is achievable, which is as follows:

$$\begin{aligned} &\Delta_t^{(i,j)} \geq 2\delta \Rightarrow S_{t,A}^{(i)} \leq S_{t,B}^{(j)} - 2\delta \text{ if true then trade, else if} \\ &\Delta_t^{(j,i)} \geq 2\delta \Rightarrow S_{t,A}^{(j)} \leq S_{t,B}^{(i)} - 2\delta \text{ is true then trade.} \end{aligned} \tag{4.88}$$

The notation in Eq. (4.88) is the same as was described in Sec. 4.3. The two conditions in Eq. (4.88) can not both be true for $i \neq j$ which intuitively means that i can sell to j and vice versa but they cannot simultaneously sell to each other; this should be apparent from how the model is constructed in Sec. 4.3 and illustrated in Fig. 4.4. Hence, if any of the conditions are true in Eq. (4.88) the mechanism comes in to play described in Eqs. (4.82)-(4.84) and visualised in the transition from Fig. 4.4 to Fig. 4.5.

4.3.4 Informed Traders

An informed trader in our model is assumed to be a trader who is a pessimistic regarding the current market price and does not believe in this market price. The informed trader believes in their price, and because of this have a particular spread that they hope to gain. Besides, when their spread is met, they do not update their price as a market maker as they do not believe in the market. In the model of trading one thus needs to define a trading trigger for this type of

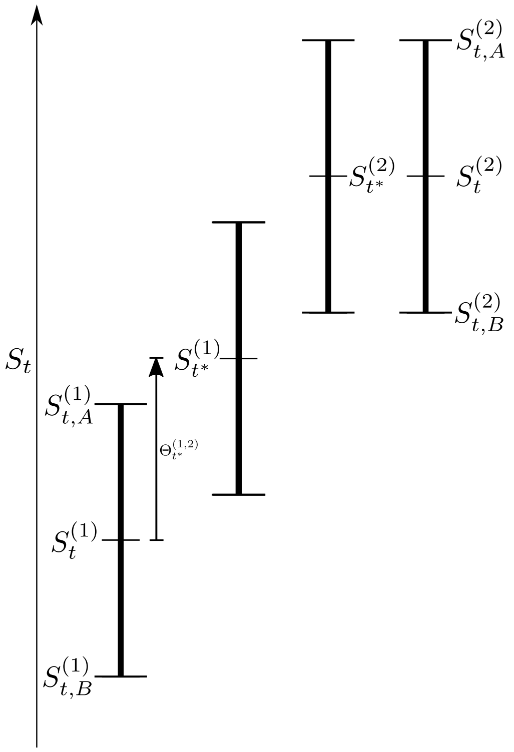


Figure 4.6: This diagram outlines the geometry in the trading mechanism between an informed trader (superscript (2)) and a market maker (superscript (1)). The parameters: half spread δ, mid-price $S_t^{(\cdot)}$, ask price $S_{t,A}^{(\cdot)}$, bid price $S_{t,B}^{(\cdot)}$, and the disagreement in bid and ask price $\Delta_t^{(\cdot,\cdot)}$. Notice that the market maker is the only agent to update his price and the informed trader does not.

trader. This is achieved as follows:

$$\begin{aligned} &\Delta_t^{(i,j)} \geq 2\delta_I \Rightarrow S_{t,A}^{(i)} \leq S_{t,B}^{(j)} - 2\delta_I \text{ if true then trade, else if} \\ &\Delta_t^{(j,i)} \geq 2\delta_I \Rightarrow S_{t,A}^{(j)} \leq S_{t,B}^{(i)} - 2\delta_I \text{ is true then trade} \end{aligned} \tag{4.89}$$

where $\delta_I \geq \delta$ is the spread of the informed trader. Hence, if one of the Eqs. (4.89) is true then a trade is a initiated and Fig. 4.6 illustrates how the prices update after a trade. From Fig. 4.6 we see that when a trade occurs and the prices of the two parties update, the only price that changes is the market maker's price; the informed trader's price remains unchanged.

4.3.5 Noise Traders

The noise trader, also known as the *liquidity trader*, is one of a class of traders that are solely looking for liquidity. By liquidity one is referring to agents that are just looking to trade (buy or sell) regardless of the price and their trigger to trade is exogenous to the market's pricing system (discussed in Sec. 4.3.3). Noise traders look to buy and sell an asset, but not simultaneously, because such traders are not looking to gain a spread. As a noise trader does not have their own spread (or perceived price) when a trade occurs with a market maker, the shift in price only happens to the market maker, this is similar to trades with an informed trader (see Fig. 4.6). The trigger that is used for this type of trader will be two independent samples at each time step from a uniform distribution with the support on the unit interval. Hence, we define the two independent random variables as $Y_B, Y_S \sim \mathcal{U}(0,1)$ which are sampled at each time step. If the samples Y_B and Y_S exceed some threshold parameter, then a trade will be triggered following:

$$\begin{aligned} &Y_B \geq \gamma \text{ if true then trade, else if} \\ &Y_S \geq \gamma \text{ is true then trade} \end{aligned} \tag{4.90}$$

where $0 \leq \gamma \leq 1$ and is a constant. The two random numbers Y_B and Y_S are used to control the buy and sell trades, respectively. Hence, if none of the inequalities in Eq. (4.90) are true then a trade is not triggered, and if both are true, then one is picked using a discrete probability distribution with equal probabilities (such as a coin flip).

4.3.6 Inventory Utility

Inventory utility (or inventory control) is introduced as a means to control the amount of trading, especially for the market maker traders. The market makers in this model will frequently trade if their spreads are small (of the size 1.0×10^{-2} and smaller) and to control this appetite for trading a utility function will be defined and implemented for each market maker. The function used for inventory utility is arbitrary, but we require it to be even, strictly positive and monotonic.

We introduce the trade-dependent spread $\delta_t^{(i)}$ where i is the index referring to a market agent and the initial value is $\delta_0^{(i)} = \delta \ \forall i$. The spread function is defined and is updated in the following way

$$\delta_{t+\Delta t}^{(i)} = \delta + \delta_e \left(\sum_{t^* \in \mathbb{T}^*} I_{t^*}^{(i)} \right)^2 \tag{4.91}$$

where Δt is the time step, t^* is the time when a trade has occurred, and $\mathbb{T}^*$ is the growing set of all trade times that have occurred up to most recent trade time t^*. The term δ_e is a constant and represents the amount the half spread adjusts after a trade and is defined as $\delta \geq \delta_e$. The buy or sell trade function $I_{t^*}^{(i)}$ is 1 when a buy trade has occurred and -1 when sell trade has occurred. The $\delta_e \left(\sum_{t^* \in \mathbb{T}^*} I_{t^*}^{(i)} \right)^2$ term is the inventory utility where the summation in the brackets is the current inventory up to the most recent trade of agent i, if no trades have occurred or the inventory is empty then this term is zero. Notice in Eq. (4.91) that we introduce the indicator function which is 1 when a trade has occurred at time t and 0 when a trade has not took place at time t; this ensures that the spread $\delta_t^{(i)}$ only changes at the time when a trade occurs.

4.4 Results

This section is a tentative investigation of the model constructed previously in Sec. 4.3 and demonstrates the numerical dynamics of the realised market price. Three market configurations (N_{MM}, N_{IT}, N_{NT}) are explored: 2 Market Makers $(2, 0, 0)$ the results of which are shown in Sec. 4.4.1, 2 Market Makers with 1 Informed Trader $(2, 1, 0)$ the results of which are shown in Sec. 4.4.3, and 2 Market Makers with 1 Noise Trader $(2, 0, 1)$ the results of which are shown in Sec. 4.4.5. Each of the market configuration has relevant parameters peculiar to the configuration, and such parameters will be varied to find features of the model. We will also consider the effect of applying inventory control, defined in Sec. 4.3.6, which will allow a closer look at the impact of the informed traders and noise traders on the market price. Also we will present the market return (price-increment) distribution and the distribution of the first passage times between successive trades.

The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$ and $X_T = 1$. The number of Monte Carlo runs changes when inventory control is and is not applied: $N_{mc} = 1.0 \times 10^5$ and $N_{mc} = 1.0 \times 10^3$ respectively for each of the parameters that are spanned over. The reason for this change in the number of Monte Carlo runs is because the number of trades between market makers is very frequent when there is no inventory control applied. This effect makes it difficult to observe the effect of adding an informed or a noise trader to the synthetic market. A summary of the parameters used throughout this results section, Sec. 4.4, is presetned in Table 4.1.

	N_{mc}	Δt	X_T	p_1	σ_i	δ_e
No-Inventory Control	1.0×10^3	1.0×10^{-6}	1	0.5	1	0
Inventory Control	1.0×10^5	1.0×10^{-6}	1	0.5	1	0.001

Table 4.1: A summary of the constant parameters used in the simulation of the trading model.

4.4.1 2 Market Makers $(2, 0, 0)$

This configuration uses a system that only consists of two market makers as explained in Sec. 4.3.3. The parameter that is spanned over is the half spread δ which is defined as the following set

$$\delta \in [0.001, 0.00105, 0.0011, 0.00115, 0.00125, 0.0015, 0.00175, 0.002, 0.003, 0.005]\,. \tag{4.92}$$

We chose this set for δ as it allows one to observe gradual changes in the number trades and market price. The a priori probabilities, as defined in Eq. (4.51), are $p_1 = p_0 = 0.5$. The distributions of the market return and first passage times, when inventory control is and is not applied with a $\delta_e = 0.001$, are shown respectively in Figs. 4.7 and 4.8. Besides, we present the moments of each of the distributions in the following tables: Table 4.2 has the estimations for the moments of the distributions for the top plots in Fig. 4.7; Table 4.3 has the estimations for the moments of the distributions for the bottom plots in Fig. 4.7; Table 4.4 has the estimations for the moments of the distributions for the top plots in Fig. 4.8; and Table 4.5 has the estimations for the moments of the distributions for the bottom plots in Fig. 4.8.

δ	# of Trades	Mean	Standard-Deviation	Skewness	Kurtosis
0.001	24205452	0.0000	0.0024	0.0269	5.4536
0.00105	2224053	0.0000	0.0025	0.0270	5.4083
0.0011	20322212	0.0000	0.0026	0.0274	5.3838
0.00115	19082652	0.0000	0.0027	0.0316	5.3384
0.00125	16730070	0.0000	0.0029	0.0307	5.2666
0.0015	11944145	0.0000	0.0034	0.0356	5.2340
0.00175	9123674	0.0001	0.0039	0.0445	5.1909
0.002	7163738	0.0001	0.0044	0.0468	5.0674
0.003	3292175	0.0001	0.0064	0.0794	5.1008
0.005	1268859	0.0004	0.0104	0.1248	5.0423

Table 4.2: We estimate the moments of the return distributions for the top plots (no inventory control) in Fig. 4.7. For the market configuration 2 Market Makers $(2,0,0)$, with no inventory control, we find the changes in the market price between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$.

δ	# of Trades	Mean	Standard-Deviation	Skewness	Kurtosis
0.001	7128514	0.0042	0.0376	1.5844	17.5113
0.00105	7095343	0.0042	0.0376	1.5822	17.5599
0.0011	7097018	0.0042	0.0376	1.5690	17.4758
0.00115	7115804	0.0042	0.0377	1.5776	17.5148
0.00125	7123717	0.0042	0.0377	1.5885	17.4934
0.0015	7157674	0.0042	0.0376	1.5830	17.3550
0.00175	7104924	0.0043	0.0378	1.5422	16.9098
0.002	7162929	0.0043	0.0378	1.5170	16.5142
0.003	7189637	0.0043	0.0380	1.4531	15.8613
0.005	7102818	0.0045	0.0386	1.2988	14.3386

Table 4.3: We estimate the moments of the return distributions for the bottom plots (inventory control) in Fig. 4.7. For the market configuration 2 Market Makers $(2,0,0)$, with inventory control, we find the changes in the market price between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^5$.

δ	Mean	Standard-Deviation	Skewness	Kurtosis
0.001	-10.8680	1.2387	-0.0837	2.7038
0.00105	-10.7953	1.2514	-0.0918	2.7022
0.0011	-10.7046	1.2554	-0.1207	2.7406
0.00115	-10.6664	1.2782	-0.1181	2.7170
0.00125	-10.5358	1.2813	-0.1290	2.7438
0.0015	-10.2289	1.3135	-0.1629	2.7762
0.00175	-9.9930	1.3532	-0.2090	2.8073
0.002	-9.7748	1.3767	-0.2254	2.8312
0.003	-9.0129	1.3992	-0.2902	2.9688
0.005	-8.1097	1.4516	-0.3247	2.9717

Table 4.4: We estimate the moments of the first passage time distributions for the top plots (no inventory control) in Fig. 4.8. For the market configuration 2 Market Makers $(2, 0, 0)$, with no inventory control, we find the changes in the time between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$.

δ	Mean	Standard-Deviation	Skewness	Kurtosis
0.001	-6.4971	2.4134	-0.2317	2.4135
0.00105	-6.4689	2.3925	-0.2295	2.4205
0.0011	-6.4456	2.3701	-0.2261	2.4361
0.00115	-6.4293	2.3499	-0.2184	2.4424
0.00125	-6.4023	2.3249	-0.2197	2.4745
0.0015	-6.3139	2.2329	-0.1945	2.5115
0.00175	-6.2378	2.1762	-0.2005	2.5732
0.002	-6.1779	2.1190	-0.2036	2.6350
0.003	-5.9657	1.9450	-0.2275	2.8322
0.005	-5.6807	1.7459	-0.3087	3.0985

Table 4.5: We estimate the moments of the first passage time distributions for the bottom plots (inventory control) in Fig. 4.8. For the market configuration 2 Market Makers $(2, 0, 0)$, with inventory control, we find the changes in the time between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^5$.

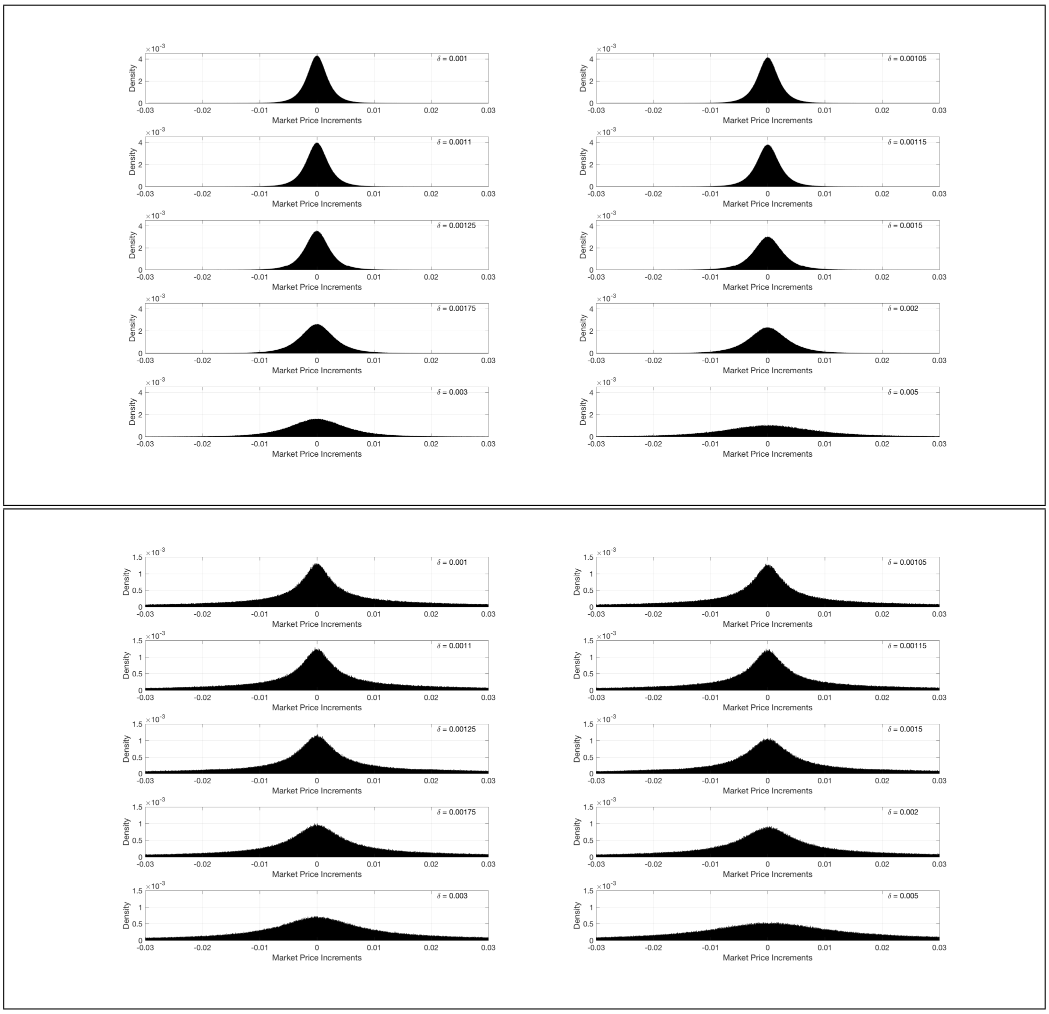


Figure 4.7: Both sets of plots (the top and bottom with black borders) are the market price increments/returns distributions that are generated by successive trades where the top plot is with no inventory control and the bottom plot is with inventory control. For the market configuration 2 Market Makers $(2, 0, 0)$ with the parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$. The number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$ (no inventory control top plot) and $N_{mc} = 1.0 \times 10^5$ (inventory control bottom plot). The y-axis is the normalised occurrences density and the x-axis is the market price increments. The spread parameter is defined as the following set δ, defined in Eq. (4.92), corresponding to the increasing value left to right and top to bottom for each of the two plots (see top right corner).

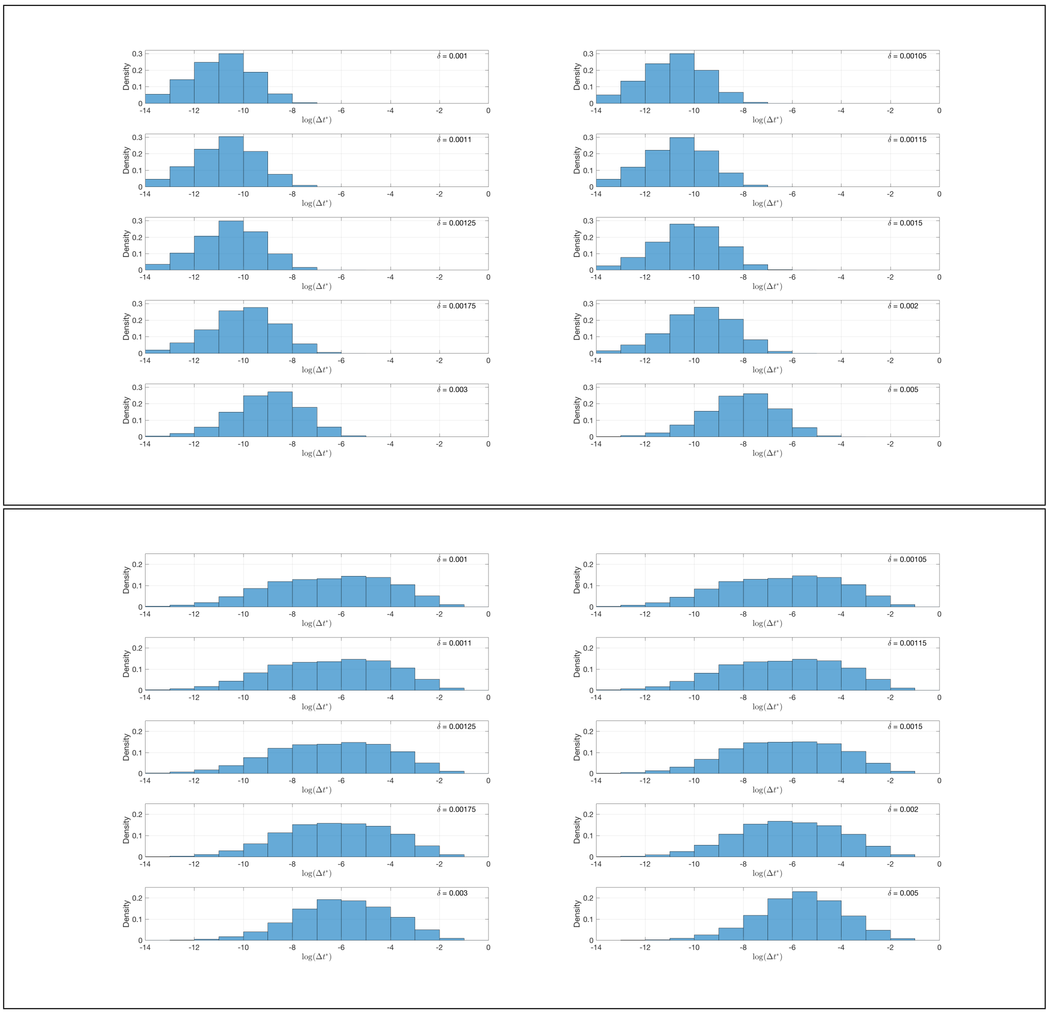


Figure 4.8: Both sets of plots (the top and bottom with black borders) are the first passage time distributions that is generated by successive trades where the top plot is with no inventory control and the bottom plot is with inventory control. For the market configuration 2 Market Makers $(2, 0, 0)$ with the parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$. The number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$ (no inventory control top plot) and $N_{mc} = 1.0 \times 10^5$ (inventory control bottom plot). The y-axis is the normalised occurrences density and the x-axis is the natural logarithm of the increments of the first passage time between trades. The spread parameter is defined as the following set δ, defined in Eq. (4.92), corresponding to the increasing value left to right and top to bottom for each of the two plots (see top right corner).

4.4.2 Discussion of $(2, 0, 0)$ Results

A market consisting of just two market makers $(2, 0, 0)$ is presented in Sec. 4.4.1. The figures Fig. 4.9 and Fig. 4.10 are examples of snap shots from a single run of the model where there is no inventory control and inventory control respectively. Note that the scale of the snap shot is not same in both Fig. 4.9 and Fig. 4.10, but is arbitrarily chosen to help give the reader a visual interpretation of the trade mechanism.

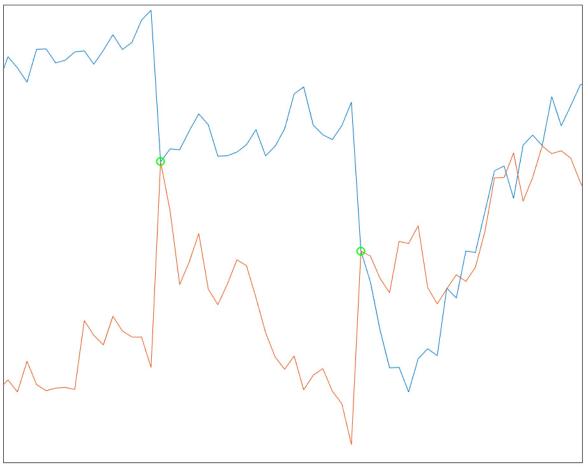


Figure 4.9: The blue and orange lines are the price processes for two market makers. The two green circles are trades between the two market makers and there is no inventory control applied.

The no inventory control results from Sec. 4.4.1 are displayed in the Tables 4.2 and 4.4 and the top of figures Fig. 4.7 (market return) and Fig. 4.8 (logarithm of time between trades). The market return is defined as the increment between two successive trades which can be seen in Fig. 4.9 and is the vertical distance between the green circles. For the market return distributions in Fig. 4.7 moving left to right and top to bottom the half spread increases

$$\delta \in [0.001, 0.00105, 0.0011, 0.00115, 0.00125, 0.0015, 0.00175, 0.002, 0.003, 0.005] .$$

We observe that the standard-deviation increases and the number of trades decrease as the half spread δ increases. The larger the spread becomes, the more the two market maker's prices have diffused away from each other, which in turn increases the average time waited for a trade. This last point is also observed in Fig. 4.8 as the distribution moves and skews more to the right.

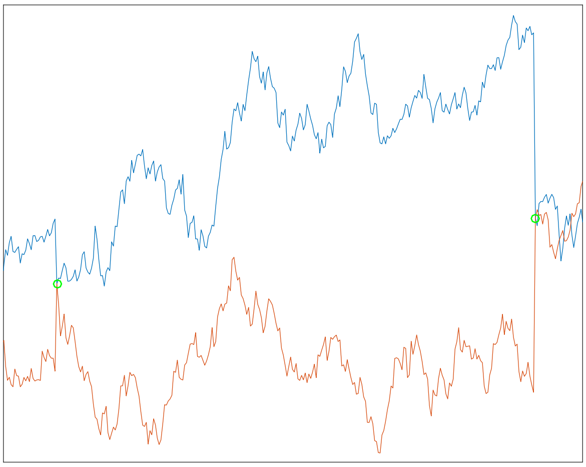


Figure 4.10: The blue and orange lines are the price process for the market makers. The green circle are the market prices created by a trade between the market makers. There is inventory control applied to both agents.

The inventory control results from Sec. 4.4.1 are displayed in the Tables 4.3 and 4.5 and the bottom figures of Fig. 4.7 (market return) and Fig. 4.8 (logarithm of time between trades). We observe that the introduction of inventory control decreases the overall number of trades, even though there are more Monte Carlo runs, shown in the Table 4.3. The number of trades is also found to be consistently around the value $\approx 7.1 \times 10^6$, which is because Eq. (4.91) is parameterised in the same way for each value of the half spread.

4.4.3 2 Market Makers and 1 Informed Trader $(2, 1, 0)$

This configuration uses a system that consists of two market makes and 1 informed trader, as explained in Sec. 4.3.4. The parameter that is spanned over is the spread of the informed trader δ_I and is defined as the following set

$$\delta_I \in [0.002, 0.003, 0.005, 0.007, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1] \,. \tag{4.93}$$

We chose this set for δ_I as the informed trader should have a larger spread than market as this agent should be in more of a disagreement, in terms of market price, before trading. The a priori probabilities, as defined in Eq. (4.51), are $p_1 = p_0 = 0.5$ and the initial spread of the market makers is set to $\delta = 0.001$. The distributions of market returns and first passage times, when inventory control is and is not applied with $\delta_e = 0.001$, are shown respectively in Figs. 4.11 and 4.12. In addition we present the moments of each of the distributions in the following tables: Table 4.6 has the estimations of the moments for the distributions for the top plots in Fig. 4.11; Table 4.7 has the estimations of the moments for the distributions for the bottom plots in Fig. 4.11; Table 4.8 has the estimations for the moments of the distributions for the top plots in Fig. 4.12; and Table 4.9 has the estimations for the moments of the distributions for the bottom plots in Fig. 4.12.

δ_I	# of Trades	Mean	Standard-Deviation	Skewness	Kurtosis
0.002	37747326	0.0000	0.0018	0.0434	16.3559
0.003	30876178	0.0000	0.0019	0.0188	4.7153
0.005	25650224	0.0000	0.0022	0.0403	5.9454
0.007	24276338	0.0000	0.0024	0.0104	4.0960
0.01	22680546	0.0000	0.0026	0.0061	4.1939
0.02	21143075	0.0000	0.0028	-0.0021	6.1621
0.03	21239750	0.0000	0.0029	0.0155	10.4071
0.05	22522936	0.0000	0.0030	0.0541	24.2690
0.075	23259898	0.0000	0.0030	0.1453	51.3784
0.1	24784621	0.0000	0.0029	0.2878	84.4465

Table 4.6: We estimate the moments of the return distributions for the top plots (no inventory control) in Fig. 4.11. For the market configuration 2 Market Makers and 1 Informed Trader $(2,1,0)$, with no inventory control, we find the changes in the market price between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$.

δ_I	# of Trades	Mean	Standard-Deviation	Skewness	Kurtosis
0.002	23392919	0.0016	0.0228	1.4827	19.9777
0.003	21826385	0.0018	0.0236	1.4021	18.5971
0.005	19354919	0.0020	0.0249	1.3030	16.3881
0.007	17591612	0.0022	0.0260	1.2342	15.2938
0.01	15504207	0.0024	0.0273	1.1480	13.9354
0.02	11469854	0.0032	0.0308	1.0075	11.2520
0.03	9404283	0.0037	0.0332	1.0006	10.3454
0.05	7413753	0.0044	0.0364	1.0581	10.2157
0.075	6418846	0.0047	0.0380	1.2068	11.1049
0.1	6045749	0.0046	0.0383	1.3465	12.4934

Table 4.7: We estimate the moments of the return distributions for the bottom plots (inventory control) in Fig. 4.11. For the market configuration 2 Market Makers and 1 Informed Trader $(2,1,0)$, with inventory control, we find the changes in the market price between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^5$.

δ_I	Mean	Standard-Deviation	Skewness	Kurtosis
0.002	-7.8246	2.3712	-0.1929	2.5306
0.003	-7.6285	2.3058	-0.2499	2.6346
0.005	-7.3373	2.2126	-0.3332	2.8192
0.007	-7.1327	2.1562	-0.3951	2.9778
0.01	-6.9087	2.1069	-0.4732	3.1996
0.02	-6.4929	2.0690	-0.6561	3.7884
0.03	-6.2728	2.0793	-0.7397	4.0735
0.05	-6.1020	2.1506	-0.7286	3.9627
0.075	-6.0909	2.2234	-0.5968	3.5147
0.1	-6.1804	2.2728	-0.4425	3.0923

Table 4.8: We estimate the moments of the first passage time distributions for the top plots (no inventory control) in Fig. 4.12. For the market configuration 2 Market Makers and 1 Informed Trader $(2, 1, 0)$, with no inventory control, we find the changes in the time between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$.

δ_I	Mean	Standard-Deviation	Skewness	Kurtosis
0.002	-11.5480	1.3992	0.0437	2.3228
0.003	-11.3310	1.4050	-0.0425	2.4182
0.005	-11.0933	1.3772	-0.1190	2.5732
0.007	-11.0423	1.3959	-0.1630	2.5756
0.01	-10.9446	1.3986	-0.2652	2.6753
0.02	-10.7959	1.3209	-0.2548	2.8206
0.03	-10.7659	1.2783	-0.1832	2.7792
0.05	-10.7935	1.2365	-0.1176	2.7558
0.075	-10.8094	1.2169	-0.0829	2.7335
0.1	-10.8754	1.2149	-0.0436	2.7081

Table 4.9: We estimate the moments of the first passage time distributions for the bottom plots (inventory control) in Fig. 4.12. For the market configuration 2 Market Makers and 1 Informed Trader $(2, 1, 0)$, with inventory control, we find the changes in the time between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^5$.

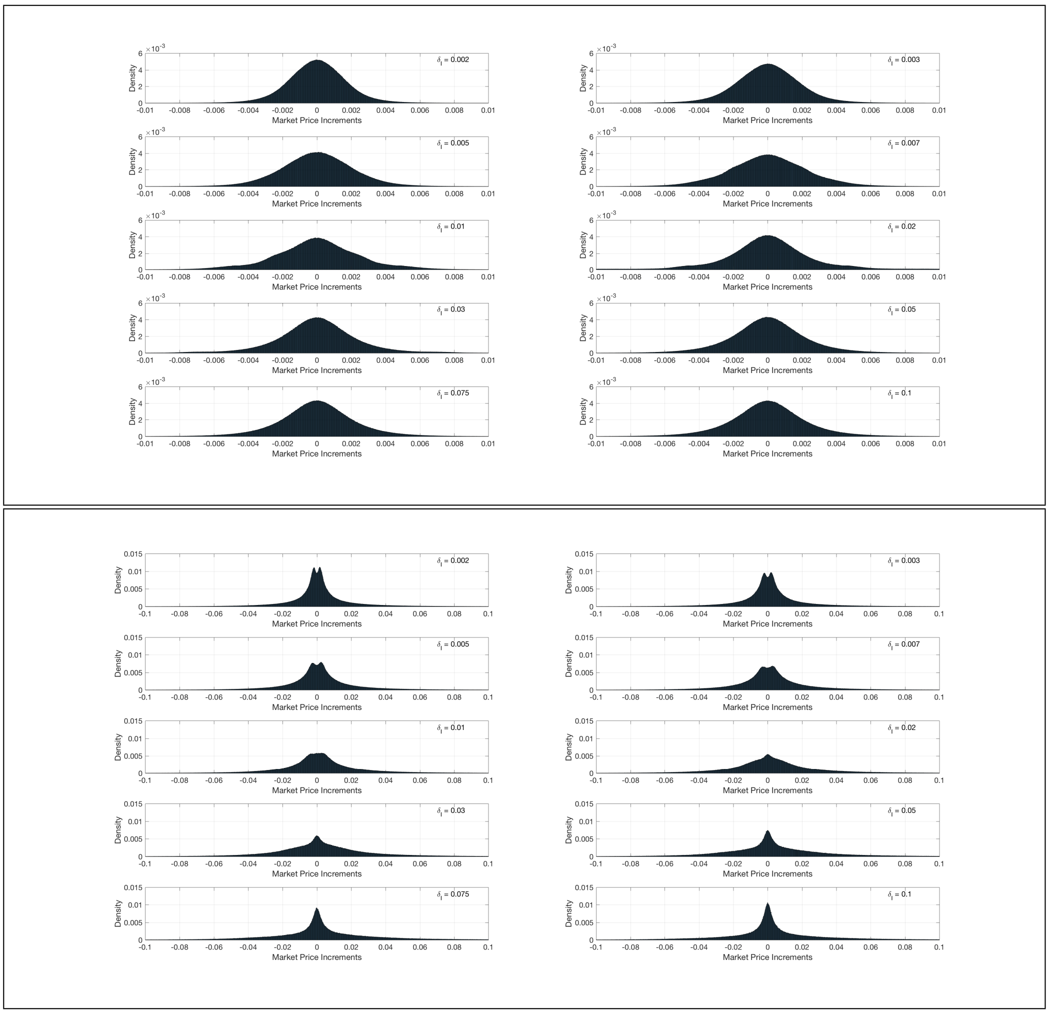


Figure 4.11: Both sets of plots (the top and bottom with black borders) are the market price increments/returns distributions that are generated by successive trades where the top plot is with no inventory control and the bottom plot is with inventory control. For the market configuration 2 Market Makers and 1 Informed Trader $(2, 1, 0)$ with the parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$. The number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$ (no inventory control top plot) and $N_{mc} = 1.0 \times 10^5$ (inventory control bottom plot). The y-axis is the normalised occurrences density and the x-axis is the market price increments. The spread parameter is defined as the following set δ_I, defined in Eq. (4.93), corresponding to the increasing value left to right and top to bottom for each of the two plots (see top right corner).

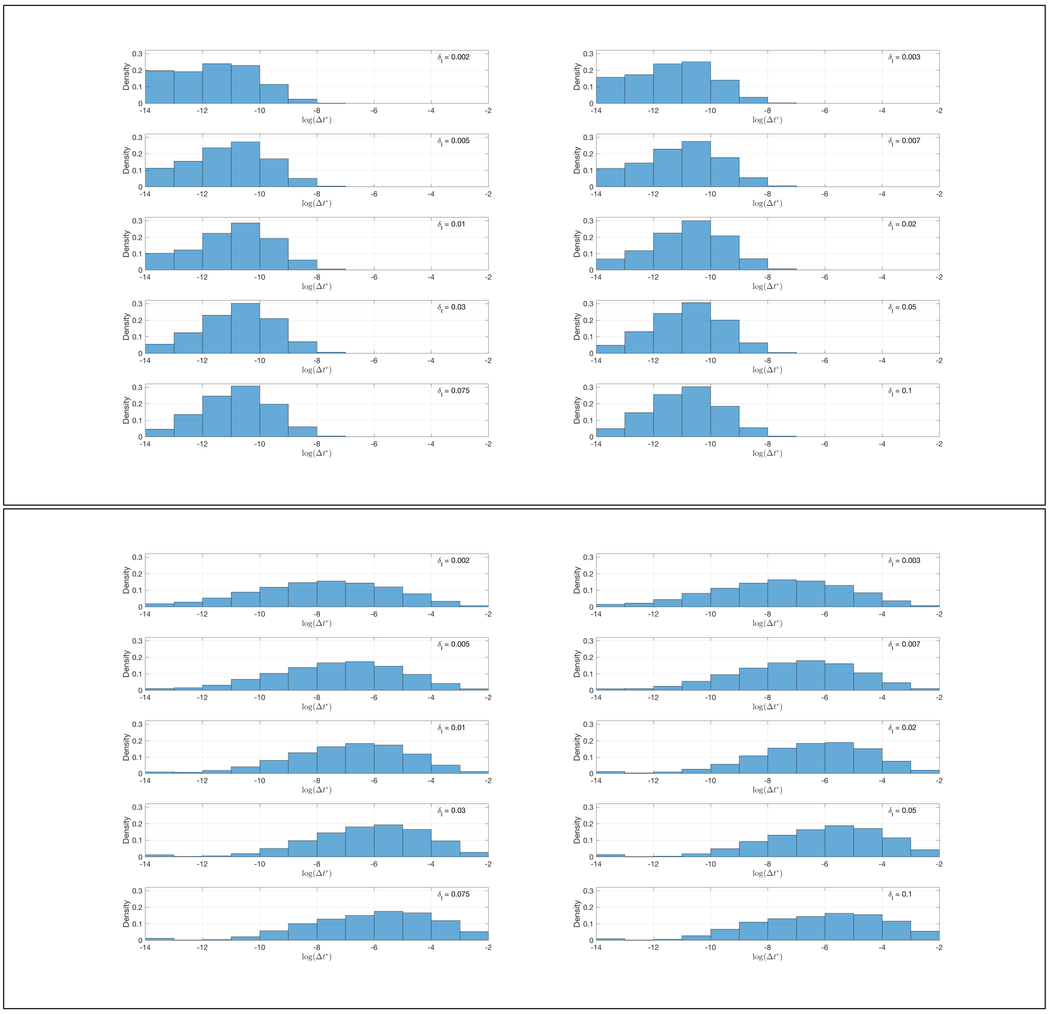


Figure 4.12: Both sets of plots (the top and bottom with black borders) are the first passage time distributions that is generated by successive trades where the top plot is with no inventory control and the bottom plot is with inventory control. For the market configuration 2 Market Makers and 1 Informed Trader $(2, 1, 0)$ with the parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$. The number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$ (no inventory control top plot) and $N_{mc} = 1.0 \times 10^5$ (inventory control bottom plot). The y-axis is the normalised occurrences density and the x-axis is the natural logarithm of the increments of the first passage time between trades. The spread parameter is defined as the following set δ_I, defined in Eq. (4.93), corresponding to the increasing value left to right and top to bottom for each of the two plots (see top right corner).

4.4.4 Discussion of $(2, 1, 0)$ Results

A market consisting of two market makers and one informed trader $(2, 1, 0)$ is presented in Sec. 4.4.3. The figures Fig. 4.13 and Fig. 4.14 are snapshots from a single run of the model where there is no inventory control and inventory control respectively. The market returns are the increments between successive market prices, referring to Fig. 4.13 the change in price is the vertical difference between adjacent circles, be they either red or green.

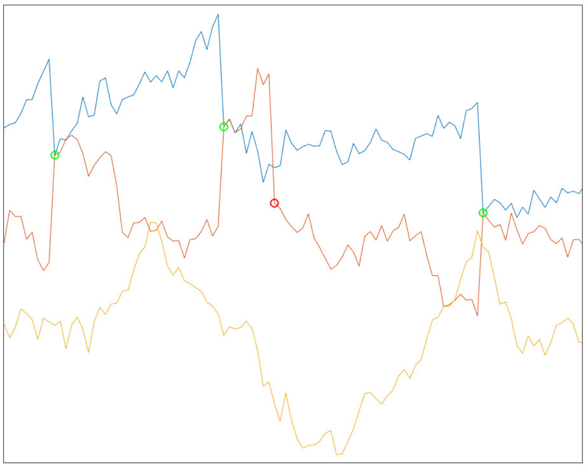


Figure 4.13: The blue and orange lines are the market maker's price processes and the yellow line is that of the informed trader. The green circle is a trade between the market makers and red circle is a trade between a market maker and the informed trader. There is no inventory control applied for any of the market agents.

The no inventory control results from Sec. 4.4.3 are displayed in the Tables 4.6 and 4.8 and the top figures of Fig. 4.11 (market return) and Fig. 4.12 (logarithm of time between trades). The market return is defined as the increment between two successive trades which can be seen in Fig. 4.13 and is the vertical distance between the green circles. For the market return distributions in Fig. 4.11 moving left to right and top to bottom the half spread of the informed trader increases

$$\delta_I \in [0.002, 0.003, 0.005, 0.007, 0.01, 0.02, 0.03, 0.05, 0.075, 0.1] \, .$$

While the market makers' spread is set to $\delta = 0.001$. We observe that the

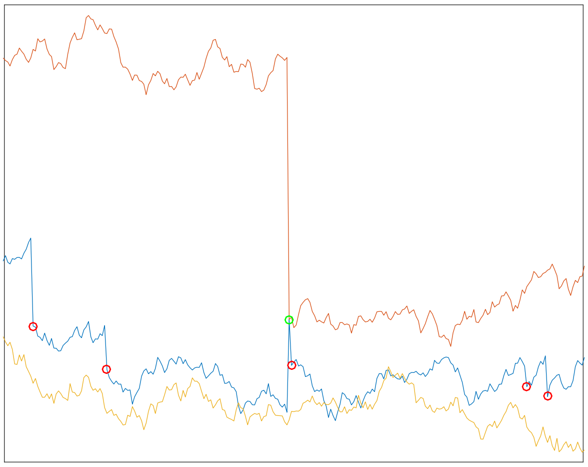


Figure 4.14: The blue and orange lines are the market maker's price processes, and the yellow line is that of the informed trader. The green circle is a trade between the market makers and the red circle is a trade between a market maker and the informed trader. There is inventory control applied for all the market agents.

standard-deviation only increases slightly (as compared to the $(2, 0, 0)$ case) and the number of trades decreases as the informed trader's half spread δ_I increases. The larger δ_I becomes, the more time the informed trader is willing to wait for the prices to diffuse away from their price, which in turn increases the average time waited to trade. This last point is also observed in Fig. 4.12 as the distribution skews more to the right. Notice, that this effect of skewing to the right in the top plot in Fig. 4.12 is not as high as seen in the corresponding plot in Fig. 4.8. The reason for this is that the two market makers do not have any inventory control and will trade very frequently as their half spread remains unchanged.

The inventory control results from Sec. 4.4.3 are displayed in the Tables 4.7 and 4.9 and the bottom figures of Fig. 4.11 (market return) and Fig. 4.12 (logarithm of time between trades). We observe that the introduction of inventory control decreases the overall number of trades which is shown in the Table 4.7.

The number of trades is also found to decrease as δ_I increases. The reason for this is that the inventory control function increases the market makers' spread. This large spread will be comparable to the informed trader's spread meaning the market makers and the informed trader will trade with each other. Another feature observed in the bottom plot in Fig. 4.12 is that the distributions are bimodal with two peaks occur at the 0.002 and −0.002. This bimodal distribution could be caused by a situation when the market makers' spread is comparable to the informed trader's spread. This, in turn, creates a clustering of trades and decreases the spread of the market makers to the minimal value but this point is a tentative one at best; this conclusion would need to be investigated further.

4.4.5 2 Market Makers and 1 Noise Trader $(2, 0, 1)$

This configuration uses a system that consists of two market makers and one noise trader, as explained in Sec. 4.3.5. The parameter that is spanned over is the threshold at which a noise trader trades γ which is defined as the following set

$$\gamma \in [0.9990, 0.9991, 0.9992, 0.9993, 0.9994, 0.9995, 0.9996, 0.9997, 0.9998, 0.9999] \,. \tag{4.94}$$

We chose this set for γ as we wanted to create a noise trader that trades a lot, which then gradual evolves in to that trades less frequently. The a priori probabilities, as defined in Eq. (4.51), are $p_1 = p_0 = 0.5$ and the initial spread of the market makers is set to $\delta = 0.001$. The distributions of increments of the market return and first passage times, when inventory control is and is not applied with $\delta_e = 0.001$, are shown respectively in Figs. 4.15 and 4.16. In addition we present the moments of each of the distributions in the following tables: Table 4.10 has the estimations for the moments of the distributions for the top plots in Fig. 4.15; Table 4.11 has the estimations for the moments of the distributions for the bottom plots in Fig. 4.15; Table 4.12 has the estimations for the moments of the distributions for the top plots in Fig. 4.16; and Table 4.13 has the estimations for the moments of the distributions for the bottom plots in Fig. 4.16.

γ	# of Trades	Mean	Standard-Deviation	Skewness	Kurtosis
0.9990	24859037	0.0000	0.0024	0.0504	8.5664
0.9991	24681375	0.0000	0.0024	0.0398	6.7691
0.9992	24419605	0.0000	0.0024	0.0271	5.6911
0.9993	24849441	0.0000	0.0024	0.0278	6.9503
0.9994	24920356	0.0000	0.0024	0.0994	6.6859
0.9995	24585881	0.0000	0.0024	0.0306	5.7095
0.9996	24722078	0.0000	0.0024	0.0302	5.6313
0.9997	24620099	0.0000	0.0024	0.0227	6.5881
0.9998	24128729	0.0000	0.0024	0.0254	5.7760
0.9999	24594965	0.0000	0.0024	0.0251	5.3597

Table 4.10: We estimate the moments of the return distributions for the top plots (no inventory control) in Fig. 4.15. For the market configuration 2 Market Makers and 1 Noise Trader $(2,0,1)$, with no inventory control, we find the changes in the market price between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$.

γ	# of Trades	Mean	Standard-Deviation	Skewness	Kurtosis
0.9990	43329183	0.0002	0.3141	0.0110	3.3466
0.9991	39366533	0.0002	0.3063	0.0123	3.4615
0.9992	35432495	0.0003	0.2975	0.0142	3.5944
0.9993	31457674	0.0004	0.2854	0.0165	3.7865
0.9994	27539006	0.0005	0.2703	0.0175	4.0390
0.9995	23618349	0.0007	0.2502	0.0247	4.3900
0.9996	19774091	0.0010	0.2230	0.0265	4.9030
0.9997	16060747	0.0015	0.1870	0.0416	5.6815
0.9998	12566303	0.0023	0.1427	0.0754	6.6753
0.9999	9378220	0.0036	0.0953	0.2129	8.0951

Table 4.11: We estimate the moments of the return distributions for the bottom plots (inventory control) in Fig. 4.15. For the market configuration 2 Market Makers and 1 Noise Trader $(2,0,1)$, with inventory control, we find the changes in the market price between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^5$.

γ	Mean	Standard-Deviation	Skewness	Kurtosis
0.9990	-7.0486	2.1021	-1.9381	6.7537
0.9991	-6.9884	2.1626	-1.9281	6.6038
0.9992	-6.9217	2.2270	-1.9155	6.4476
0.9993	-6.8469	2.2979	-1.8957	6.2662
0.9994	-6.7640	2.3726	-1.8711	6.0788
0.9995	-6.6739	2.4603	-1.8341	5.8437
0.9996	-6.5653	2.5471	-1.7897	5.6188
0.9997	-6.4383	2.6174	-1.7283	5.4080
0.9998	-6.2839	2.6467	-1.6287	5.2060
0.9999	-6.1386	2.5727	-1.3458	4.7420

Table 4.12: We estimate the moments of the first passage time distributions for the top plots (no inventory control) in Fig. 4.16. For the market configuration 2 Market Makers and 1 Noise Trader $(2, 0, 1)$, with no inventory control, we find the changes in the time between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$.

γ	Mean	Standard-Deviation	Skewness	Kurtosis
0.9990	-10.8926	1.2560	-0.0916	2.7516
0.9991	-10.8874	1.2570	-0.0866	2.7412
0.9992	-10.8680	1.2488	-0.1020	2.7714
0.9993	-10.9016	1.2594	-0.0801	2.7387
0.9994	-10.8922	1.2460	-0.0760	2.7543
0.9995	-10.8824	1.2471	-0.0766	2.7503
0.9996	-10.8923	1.2510	-0.0819	2.7504
0.9997	-10.8916	1.2504	-0.0738	2.7411
0.9998	-10.8514	1.2276	-0.0752	2.7725
0.9999	-10.8763	1.2315	-0.0677	2.7377

Table 4.13: We estimate the moments of the first passage time distributions for the bottom plots (inventory control) in Fig. 4.16. For the market configuration 2 Market Makers and 1 Noise Trader $(2, 0, 1)$, with inventory control, we find the changes in the time between successive trades and present the sample statistics. The parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$, and the number of Monte Carlo runs $N_{mc} = 1.0 \times 10^5$.

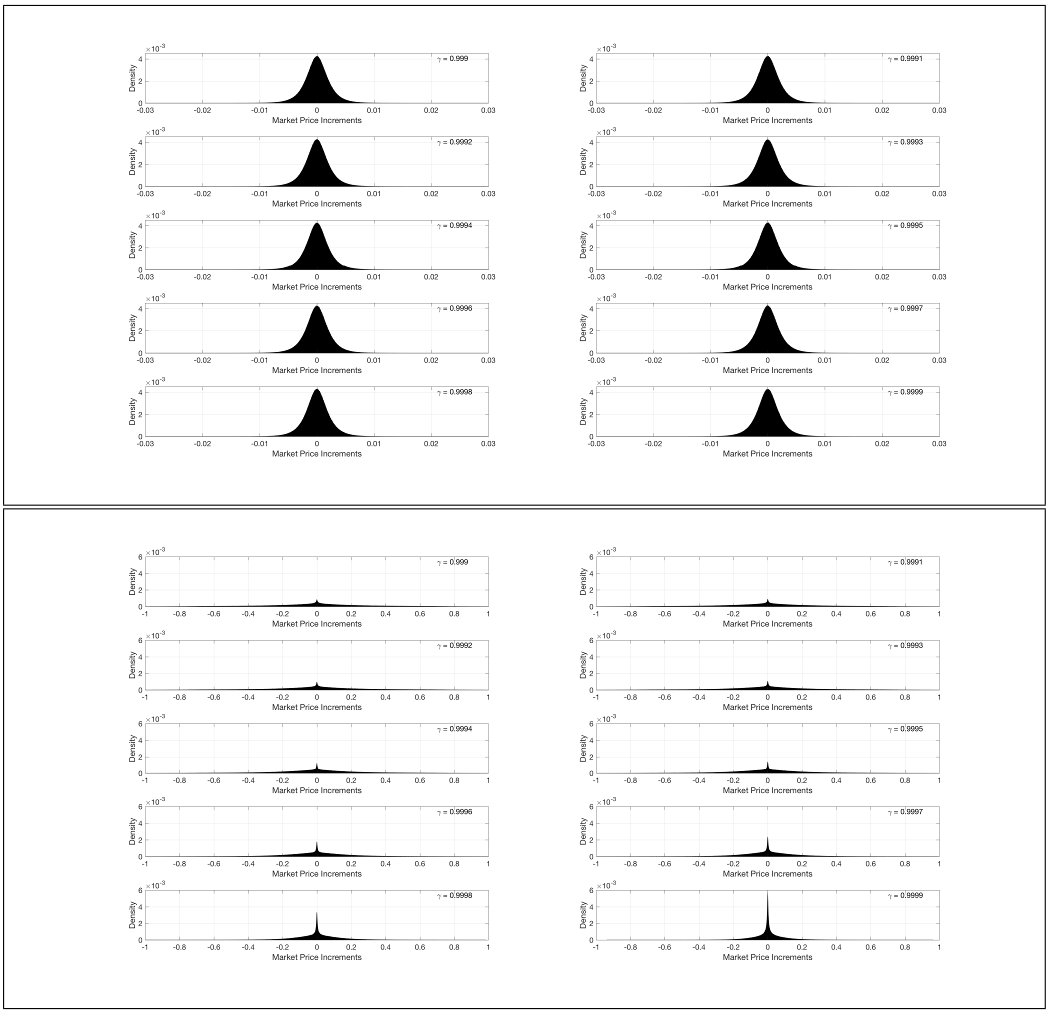


Figure 4.15: Both sets of plots (the top and bottom with black borders) are the market price increments/returns distributions that are generated by successive trades where the top plot is with no inventory control and the bottom plot is with inventory control. For the market configuration 2 Market Makers and 1 Noise Trader $(2, 0, 1)$ with the parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$. The number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$ (no inventory control top plot) and $N_{mc} = 1.0 \times 10^5$ (inventory control bottom plot). The y-axis is the normalised occurrences density and the x-axis is the market price increments. The spread parameter is defined as the following set γ, defined in Eq. (4.94), corresponding to the increasing value left to right and top to bottom for each of the two plots (see top right corner).

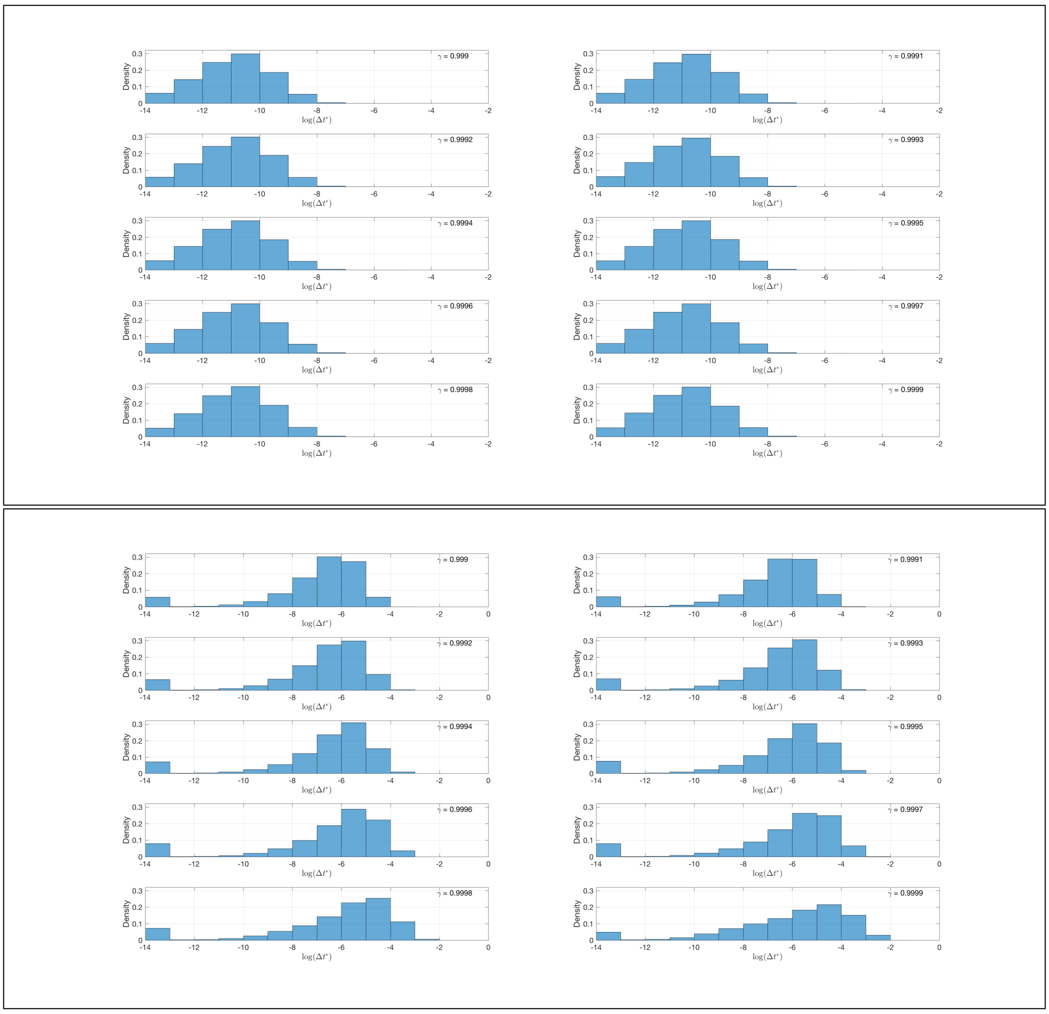


Figure 4.16: Both sets of plots (the top and bottom with black borders) are the first passage time distributions that is generated by successive trades where the top plot is with no inventory control and the bottom plot is with inventory control. For the market configuration 2 Market Makers and 1 Noise Trader $(2,0,1)$ with the parameters that are unchanged for all simulations are as follows: the time step $\Delta t = 1.0 \times 10^{-6}$, $X_T = 1$. The number of Monte Carlo runs $N_{mc} = 1.0 \times 10^3$ (no inventory control top plot) and $N_{mc} = 1.0 \times 10^5$ (inventory control bottom plot). The y-axis is the normalised occurrences density and the x-axis is the natural logarithm of the increments of the first passage time between trades. The spread parameter is defined as the following set γ, defined in Eq. (4.94), corresponding to the increasing value left to right and top to bottom for each of the two plots (see top right corner).

4.4.6 Discussion of $(2, 0, 1)$ Results

A market consisting of two market makers and one noise trader $(2, 1, 0)$ is presented in Sec. 4.4.5. The figures Fig. 4.17 and Fig. 4.18 are snapshots from a single run of the model where there is no inventory control and inventory control respectively. The market returns are the increments between successive market prices, referring to Fig. 4.17 the change in price is the vertical difference between adjacent circles, be them either red or green.

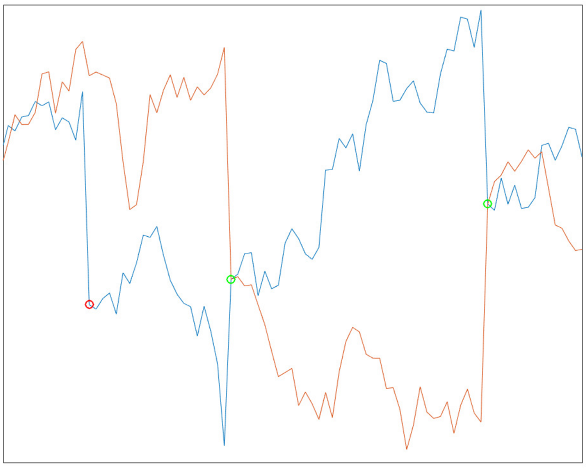


Figure 4.17: The blue and orange lines are the market makers' price processes. The green circle is a trade between the market makers, and the red circle is a trade between a market maker and noise trader. There is no inventory control applied for the market agents.

The no inventory control results from Sec. 4.4.5 are displayed in the Tables 4.10 and 4.12 and the top of figures Fig. 4.15 (market return) and Fig. 4.16 (log time between trades). The market return is defined as the increment between two successive trades which can be seen in Fig. 4.17 to be the vertical distance between the adjacent circles. The parameter that is spanned over is the threshold

γ which is defined as the following set

$$\gamma \in [0.9990, 0.9991, 0.9992, 0.9993, 0.9994, 0.9995, 0.9996, 0.9997, 0.9998, 0.9999]$$

While the market makers' spread is set to $\delta = 0.001$. We observe that the standard-deviation does not change and is of similar magnitude as compared to the $(2, 0, 0)$ and $(2, 1, 0)$ cases. The number of trades remains reasonably constant when the noise trader's parameter γ increases. These two observations and with the addition of the top plot in Figs. 4.15 and 4.16 we can see the values of the γ parameter has little to no effect on the distribution of market returns. This result could be alleviated if the half spread of the market makers is increased (or inventory control applied) which will make the effect of noise trader more apparent.

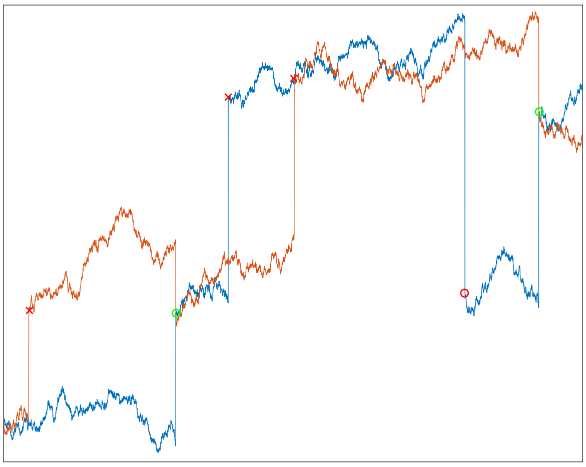


Figure 4.18: The blue and orange lines are the market makers' price processes. The green circle is a trade between the market makers and the red circle is a trade between a market maker and a noise trader. There is inventory control applied for all the market agents apart from the noise trader. Notice that there is no price trace for the noise trader which is in line with the model construction.

The inventory control results from Sec. 4.4.5 are displayed is the Tables

4.11 and 4.13 and the bottom figures of Fig. 4.15 (market return) and Fig. 4.16 (logarithm of time between trades). We observe that the introduction of inventory control increases the overall standard deviation of the market return distribution shown in Table 4.11. The number of trades is also found to decrease as γ increases. The reason that the number of trades decreases when γ increases is that the probability that the trade trigger Eq. (4.90) is activated is decreasing. Another feature that we see is that no matter how big the market makers' spread becomes the noise trader will trade when Eq. (4.90) is activated and this causes swings in the market price. This effect of price swinging is most apparent in our model parameters when $\gamma = 0.9990$ and the least when $\gamma = 0.9999$. We see that in the bottom plot in Fig. 4.16 that there is a small peak in the likelihood of trading at the smallest time step. We speculate that this peak emerges from the initial trading between market makers at times close to $t = 0$. The spread grows from $\delta = 0.001$ until it becomes too big for market makers to trade with each other, and only the noise trader can trade. This final observation and explanation is a tentative one and would need further investigation.

4.5 Non-linear Information Rates and their Limits

We return to a point made in Sec. 4.1.2, that is that the true information part $\sigma X_T t$ of the information process is arbitrary chosen by the modeller (for mathematical convenience). This is a somewhat unsatisfactory feature of the model, which will be investigated in this section and Sec. 4.6. This section explores the arbitrariness of the BHM model and we start by redefining the information process Eq. (4.11) as

$$\xi_t \triangleq f(t) X_T + \beta_{tT} \tag{4.95}$$

where the BHM model is recovered by setting $f(t) = \sigma t$. This linear assumption is unfounded by economic or financial principles. The modeller in the BHM framework is free to choose $f(t)$ as long as in the domain $t \in [0, T]$ it is smooth and continuous, which ensures that the process Eq. (4.95) is Markov [130, 131]. This section also aims to examine the limits of the BHM model when $f(t)$ is not linear and when we have correlated price processes.

We proceed with a thought experiment to test the applicability of the BHM model and this experiment will be based on a two horse race market. Such markets were discussed in Chapters 2 and 3. As before, in Eq. (4.2), we define the random variable $X_T \in \{0,1\}$ which represents the final outcome of an uncertain terminal (at time $t = T$) event. This random variable X_T is theorised to represent the outcome of a horse race. It is assumed that the race finishes at a fixed time $t = T$ which in reality is not strictly true as a race finishes when the course distance is completed*. The market consists of one gambler and the role of the gambler is to try and work out which horse is going to win the race while the race is in-play. Consider a race that consists of two horses, denoted as horse 1 and 2, and assuming the gambler has access to the following information processes

$$\begin{pmatrix} \xi_t^{(1)} \\ \xi_t^{(2)} \end{pmatrix} = \begin{pmatrix} f(t) \\ g(t) \end{pmatrix} X_T + \begin{pmatrix} \beta_{tT}^{(1)} \\ \beta_{tT}^{(2)} \end{pmatrix} \tag{4.96}$$

where X_T indicates which horse is the winning horse. In reality at $t = 0$ both horses can win the race, hence the final state of the system can either be $X_T^{(1)} = 1$ and $X_T^{(2)} = 0$ or $X_T^{(1)} = 0$ and $X_T^{(2)} = 1$: which applies it is not known until $t = T$. The BHM model can not capture this behaviour and does not make the claim that it is trying to predict the final state of X_T, only model how agents process the information in determining a price of a future cash-flow. To ensure the model is now consistent with our narrative, that is the information process will converge to 1 for the winner and zero otherwise, we will require the functions $f(t)$ and $g(t)$ to converge to the appropriate values such that $\lim_{t\to T} f(t) = 1$ and $\lim_{t\to T} g(t) = 0$ or $\lim_{t\to T} f(t) = 0$ and $\lim_{t\to T} g(t) = 1$. The random variable X_T will be set to one and the interpretation of this is that X_T indicates that both horse can win at $t < T$ but the functions $f(t)$ and $g(t)$ will determine the actual winner at $t = T$.

*One can also take horse gambling price data and standardise the race time T to unity such that $t \in [0,1]$.

Considering $f(t)$ and $g(t)$ to be of quadratic forms the Eq. (4.96) becomes

$$\begin{pmatrix} \xi_{tT}^{(1)} \\ \xi_{tT}^{(2)} \end{pmatrix} = \begin{pmatrix} (t/T)^2 \\ -(t/T)^2 + t/T \end{pmatrix} \sigma X_T + \begin{pmatrix} \beta_{tT}^{(1)} \\ \beta_{tT}^{(2)} \end{pmatrix} . \tag{4.97}$$

The final time can be set to $T = 1$ which make the models notation more stream lined

$$\begin{pmatrix} \xi_{t}^{(1)} \\ \xi_{t}^{(2)} \end{pmatrix} = \begin{pmatrix} t^2 \\ t(1-t) \end{pmatrix} \sigma X_T + \begin{pmatrix} \beta_{t}^{(1)} \\ \beta_{t}^{(2)} \end{pmatrix} . \tag{4.98}$$

The moments of the processes are found to be $\mathbb{E}\left[\xi_t^{(1)}\right] + \mathbb{E}\left[\xi_t^{(2)}\right] = \sigma x t$ and $\mathbb{V}\left[\xi_t^{(1)}\right] = \mathbb{V}\left[\xi_t^{(2)}\right] = t(1-t)$; with the covariance $\mathbb{C}\left[\xi_t^{(1)}, \xi_t^{(2)}\right] = \mathbb{E}\left[\beta_t^{(1)}\beta_t^{(2)}\right] = 0$ and if we define $0 \le s < t \le 1$, the auto-covariance is $\mathbb{C}\left[\xi_s^{(1)}, \xi_t^{(2)}\right] = \mathbb{E}\left[\beta_s^{(1)}\beta_t^{(2)}\right] = s(1-t)$. Combining all of these results and using the definition on page 111 of the H. M. Mahmoud book [132] one finds that the expectation vector and covariance matrix of Eq. (4.98) can be written as

$$\bar{\mu} = \begin{pmatrix} \sigma x s^2 \\ \sigma x t(1-t) \end{pmatrix} \quad \text{and} \quad \bar{\bar{\Sigma}} = \begin{pmatrix} s(1-s) & s(1-t) \\ s(1-t) & t(1-t) \end{pmatrix} . \tag{4.99}$$

The bivariate normal distribution of the two information processes $\xi_s^{(1)}$ and $\xi_t^{(2)}$ is thus $\rho\left(\xi_s^{(1)}, \xi_t^{(2)} \mid X_T = x\right) \sim \mathcal{N}\left(\xi_s^{(1)}, \xi_t^{(2)} \mid \bar{\mu}, \bar{\bar{\Sigma}}\right)$, which explicitly is written as

$$\begin{aligned} \rho\left(\xi_s^{(1)}, \xi_t^{(2)} \mid X_T = x\right) &= \frac{1}{2\pi\sqrt{\det\left(\bar{\bar{\Sigma}}\right)}} \exp\left(-\tfrac{1}{2}\left(\bar{\xi} - \bar{\mu}\right)' \bar{\bar{\Sigma}}^{-1}\left(\bar{\xi} - \bar{\mu}\right)\right) \\ &= \frac{\exp\left(-\frac{1}{2s(t-s)(1-t)}\left(s(1-s)\left(\xi_s^{(1)} - xs^2\right)^2 + t(1-t)\left(\xi_t^{(2)} - xt(1-t)\right)^2 - 2s(1-t)\xi_s^{(1)}\xi_t^{(2)}\right)\right)}{2\pi\sqrt{s(t-s)(1-t)}} , \end{aligned} \tag{4.100}$$

where the vector $\bar{\xi}$ is defined as $\bar{\xi} = \left(\xi_s^{(1)}, \xi_t^{(2)}\right)'$ and the dash denotes the transpose. It can be seen that if the times are $s = t$, the distribution Eq. (4.100) collapses to $\lim_{s \to t} \rho\left(\xi_s^{(1)}, \xi_t^{(2)} \mid X_T = x\right) \to \infty$. To avoid such a situation one must make $\xi_t^{(1)} \perp\!\!\!\perp \xi_t^{(2)} \Rightarrow \rho\left(\xi_t^{(1)}, \xi_t^{(2)} \mid X_T = x\right) = \rho\left(\xi_t^{(1)} \mid X_T = x\right)\rho\left(\xi_t^{(2)} \mid X_T = x\right)$ and $\mathbb{C}\left[\xi_t^{(1)}, \xi_t^{(2)}\right] =$

$\mathbb{E}\left[\beta_t^{(1)}\beta_t^{(2)}\right] = 0$, hence Eq. (4.100) becomes

$$\rho\left(\xi_t^{(1)}, \xi_t^{(2)} \mid X_T = x\right) = \tfrac{1}{2\pi t(1-t)} \exp\left(-\tfrac{\left(\xi_t^{(1)} - \sigma x t^2\right)^2}{2t(1-t)}\right) \exp\left(-\tfrac{\left(\xi_t^{(2)} - \sigma x t(1-t)\right)^2}{2t(1-t)}\right) . \tag{4.101}$$

From a theoretical point-of-view Eq. (4.101) is rather a limiting one as one would like to couple the information process Eq. (4.98) with an interaction term $\xi_t^{(1)}\xi_t^{(2)}$, but Eq. (4.100) makes this impossible because the two information process are independent $\Rightarrow \rho\left(\xi_t^{(1)}, \xi_t^{(2)} \mid X_T = x\right) = \rho\left(\xi_t^{(1)} \mid X_T = x\right)\rho\left(\xi_t^{(2)} \mid X_T = x\right)$. Not having such an interaction term leads to an inconsistency with the argument and we will try to highlight this limitation of the model. The univariate components of Eq. (4.100) are found by marginalisation* giving

$$\begin{aligned} \rho\left(\xi_t^{(1)} \mid X_T^{(1)} = x\right) &= \sqrt{\frac{1}{2\pi t(1-t)}} \exp\left(-\frac{1}{2t(1-t)}\left(\xi_t^{(1)} - \sigma x t^2\right)^2\right) \\ \rho\left(\xi_t^{(2)} \mid X_T^{(2)} = x\right) &= \sqrt{\frac{1}{2\pi t(1-t)}} \exp\left(-\frac{1}{2t(1-t)}\left(\xi_t^{(2)} - \sigma x t(1-t)\right)^2\right) \end{aligned} \tag{4.102}$$

after expanding out the exponent one finds that

$$\begin{aligned} \rho\left(\xi_t^{(1)} \mid X_T^{(1)} = x\right) &= \mathcal{N}\left(\xi_t^{(1)} \mid 0, t(1-t)\right) \exp\left(\tfrac{\sigma x t}{(1-t)}\left(\xi_t^{(1)} - \tfrac{1}{2}\sigma x t^2\right)\right) \\ \rho\left(\xi_t^{(2)} \mid X_T^{(2)} = x\right) &= \mathcal{N}\left(\xi_t^{(2)} \mid 0, t(1-t)\right) \exp\left(\sigma x\left(\xi_t^{(2)} - \frac{1}{2}\sigma x t(1-t)\right)\right) . \end{aligned} \tag{4.103}$$

Using Bayes' theorem, as in Eq. (4.22), one finds that the conditional densities that are needed to calculated the price (or implied back probabilities, see Eq. (2.1)) of horses 1 and 2 are

$$\begin{aligned} \rho\left(x \mid \xi_t^{(1)}\right) &= \frac{f(x)\exp\left(\frac{t}{(1-t)}\sigma x\left(\xi_t^{(1)} - \frac{1}{2}\sigma x t^2\right)\right)}{\int_0^1 f(z)\exp\left(\frac{t}{(1-t)}\sigma z\left(\xi_t^{(1)} - \frac{1}{2}\sigma z t^2\right)\right)\mathrm{d}z} \to \phi_{it}^{(1)} = \frac{p_i \exp\left(\frac{t}{(1-t)}\sigma x_i\left(\xi_t^{(1)} - \frac{1}{2}\sigma x_i t^2\right)\right)}{\Sigma_0^1 p_i \exp\left(\frac{t}{(1-t)}\sigma x_i\left(\xi_t^{(1)} - \frac{1}{2}\sigma x_i t^2\right)\right)} \\ \rho\left(x \mid \xi_t^{(2)}\right) &= \frac{f(x)\exp\left(\sigma x\left(\xi_t^{(2)} - \frac{1}{2}\sigma x t(1-t)\right)\right)}{\int_0^1 f(z)\exp\left(\sigma z\left(\xi_t^{(2)} - \frac{1}{2}\sigma z t(1-t)\right)\right)\mathrm{d}z} \to \phi_{it}^{(2)} = \frac{p_i \exp\left(\sigma x_i\left(\xi_t^{(2)} - \frac{1}{2}\sigma x_i t(1-t)\right)\right)}{\Sigma_0^1 p_i \exp\left(\sigma x_i\left(\xi_t^{(2)} - \frac{1}{2}\sigma x_i t(1-t)\right)\right)} . \end{aligned} \tag{4.104}$$

*By marginalisation one means the following procedure $\rho\left(\xi_s^{(1)} \mid X_T = x\right) = \int \mathrm{d}\xi_t^{(2)} \rho\left(\xi_s^{(1)}, \xi_t^{(2)} \mid X_T = x\right)$ and $\rho\left(\xi_t^{(2)} \mid X_T = x\right) = \int \mathrm{d}\xi_t^{(1)} \rho\left(\xi_s^{(1)}, \xi_t^{(2)} \mid X_T = x\right)$.

The right arrow is indicating that x is transformed to a discrete state, which in this case is a binary variable. The price processes are found in the same way as in Sec. 4.2.3 giving

$$\begin{aligned}\phi_{1t}^{(1)} &= \left(1 + \tfrac{p_0}{p_1} \exp\left(-\tfrac{t}{1-t}\sigma\left(\xi_t^{(1)} - \tfrac{1}{2}\sigma t^2\right)\right)\right)^{-1} \\ \phi_{1t}^{(2)} &= \left(1 + \tfrac{p_0}{p_1} \exp\left(-\sigma\left(\xi_t^{(2)} - \tfrac{1}{2}\sigma t(1-t)\right)\right)\right)^{-1} .\end{aligned} \tag{4.105}$$

One finds a few inconsistencies from Eq. (4.105): the first point is that the two prices do not add up to unity $\phi_{1t}^{(1)} + \phi_{1t}^{(2)} \neq 1$, which could be brought about by the fact that there is no interaction term between the two information process because $\xi_t^{(1)} \perp\!\!\!\perp \xi_t^{(2)}$; and the second point is that $\lim_{t\to 1} \phi_{1t}^{(2)} \neq 1$. This second point is a important observation because as was shown in Eq. (4.70) that this condition must be true, hence it must be $\lim_{t\to 1} \phi_{1t}^{(2)} = 1$ for the model to be consistent. This indicates that there are restriction on what $f(t)$ and $g(t)$ can be such that the following condition for both functions

$$\lim_{t\to T} \frac{f(t)}{t(t-1)} = \lim_{t\to T} \frac{f(t)}{\mathbb{V}\left[\beta_{tT}\right]} \to \infty \tag{4.106}$$

is true. This requirement ensures that the conditional density has the limiting property $\lim_{t\to T} \phi_{it} = \mathbb{1}_{\{X_T = x_i\}}$ for $i = 0, 1$; which means that the price will have the correct terminal value, that is $\lim_{t\to T} S_t = X_T$.

4.6 Generalisation of Non-linear Information Rate

This section has the objective to create a class of information processes that are not linear as in the BHM model Eq. (4.11), which converge in the same way as Eq. (4.106) or more generally Eq. (4.70). This reformulation of Eq. (4.11) is important when fitting to real data, which we will return to in Sec. 4.7.2. Consider an information process of the form

$$\xi_t = g(t) X_T + \beta_{tT} \tag{4.107}$$

where the conditional density is $\rho(\{\xi_t \mid X_T = x\}) \sim \mathcal{N}(g(t)x, t(T-t)/T)$ and we assume that $\lim_{t\to T} \frac{g(t)}{\mathbb{V}[\beta_{tT}]} \to \infty$. To find if Eq. (4.107) is Markovian consider a sequence of increasing times (but decreasing index) $0 < u_n < u_{n-1} < \cdots < u_2 < u_1 < t \leq T$. Define the following sequence

$$\kappa_i \triangleq \frac{\beta_{u_i T}}{g(u_i)} - \frac{\beta_{u_{i+1}T}}{g(u_{i+1})} = \frac{\xi_{u_i}}{g(u_i)} - \frac{\xi_{u_{i+1}}}{g(u_{i+1})} \tag{4.108}$$

where $i = 1, 2, \ldots, n$. The covariance between κ_i and β_{tT} decreases as $n \to \infty$. This can be shown by taking the following expectation in the limit $n \to \infty$ such that

$$\lim_{n\to\infty} \left\{\mathbb{E}^{\mathbb{Q}}\left[\beta_{tT}\kappa_i\right]\right\} = \lim_{n\to\infty} \left\{\frac{u_i(T-t)}{g(u_i)T} - \frac{u_{i+1}(T-t)}{g(u_{i+1})T}\right\} \to 0\ , \tag{4.109}$$

therefore Eq. (4.107) is Markovian in the continuous limit. The dynamical process of the non-linear information is found to be

$$\begin{aligned} \mathrm{d}\xi_t &= \mathrm{d}\left[g(t)X_T + \beta_{tT}\right] \\ &= \frac{\partial g}{\partial t} X_T \mathrm{d}t - \frac{\beta_{tT}}{T-t}\mathrm{d}t + \mathrm{d}B_t \\ &= \frac{1}{T-t}\left(\frac{\partial g}{\partial t} X_T (T-t) - \beta_{tT}\right)\mathrm{d}t + \mathrm{d}B_t \end{aligned} \tag{4.110}$$

which is semimartingale if and only if $g(.)$ is a function with bound variation. The quadratic variation of Eq. (4.107) is thus found to be

$$\mathbb{E}\left[\mathrm{d}\xi_t^2\right] = \mathbb{E}\left[\left(\left(X_T\frac{\partial g}{\partial t} - \frac{B_T}{T}\right)\mathrm{d}t + \mathrm{d}B_t\right)^2\right] = \mathrm{d}t \tag{4.111}$$

which is a necessary condition for application of Itô's lemma. Applying Bayes' theorem, as in Eq. (4.24), and defining $\rho(\xi_t \mid X_T = x) \triangleq \phi_t(x)$ to streamline notation, one finds that

$$\phi_t(x) = \frac{f(x)\exp\left(\frac{T}{t(T-t)}\hat{g}_t x(y - \frac{1}{2}\hat{g}_t x)\right)}{\int_0^1 f(z)\exp\left(\frac{T}{t(T-t)}\hat{g}_t z(y - \frac{1}{2}\hat{g}_t z)\right)\mathrm{d}z} \tag{4.112}$$

where is $f(x)$ is the a priori probability density function of $X_T = x$ and $x \in [0, 1]$. A hat has been placed on the non-linear information rate $\hat{g}_t$ as an indication that this will be estimated from data, which will be done in Sec. 4.7. The price is found in the same way as in Eq. (4.25) where price S_t is equivalent to the function $S(\xi_t, t)$, which will be the function we apply Itô's lemma to. Hence, Itô's lemma applied to the price function $S(\xi_t, t)$ is as follows

$$\begin{aligned}
\mathrm{d}S(\xi_t,t) &= \frac{\partial S_t}{\partial y}\mathrm{d}\xi_t + \frac{\partial S_t}{\partial t}\mathrm{d}t + \frac{1}{2}\frac{\partial^2 S_t}{\partial y^2}\mathrm{d}\xi_t^2 \\
&= \left(\int_0^1 x\tfrac{\partial}{\partial y}\phi_t(x)\mathrm{d}x\right)\mathrm{d}\xi_t + \left(\int_0^1 x\tfrac{\partial}{\partial t}\phi_t(x)\mathrm{d}x\right)\mathrm{d}t + \left(\tfrac{1}{2}\int_0^1 x\frac{\partial^2}{\partial y^2}\phi_t(x)\mathrm{d}x\right)\mathrm{d}\xi_t^2 \\
&= \left(\int_0^1 x\Gamma_1(\xi_t,t)\mathrm{d}x\right)\mathrm{d}\xi_t + \left(\int_0^1 x\Gamma_2(\xi_t,t)\mathrm{d}x + \frac{1}{2}\int_0^1 x\Gamma_3(\xi_t,t)\mathrm{d}x\right)\mathrm{d}t
\end{aligned} \tag{4.113}$$

where the functions $\Gamma_1(\xi_t,t)$, $\Gamma_2(\xi_t,t)$ and $\Gamma_3(\xi_t,t)$ are defined as the partial differential terms in the integrands. The aim is to calculate each of these three functions and plug them into Eq. (4.113), which will lead to a dynamic expression of the price process. To reduce the notation in the calculation, define the two functions

$$\begin{aligned}
A &\triangleq f(x)\exp\left(\frac{T}{t(T-t)}\hat{g}_t x(\xi_t - \tfrac{1}{2}\hat{g}_t x)\right) \\
B &\triangleq \int_0^1 f(x)\exp\left(\frac{T}{t(T-t)}\hat{g}_t x(\xi_t - \tfrac{1}{2}\hat{g}_t x)\right)\mathrm{d}x \ ,
\end{aligned} \tag{4.114}$$

such that a third function can be defined in terms of A and B

$$C \triangleq \int_0^1 xA\mathrm{d}x = B\mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t\right] = BS(\xi_t,t) \ . \tag{4.115}$$

The functions defined in Eq. (4.114) naturally lead to the useful expression for Eq. (4.112) which is $\phi_t(x) = \frac{A}{B}$. The $\Gamma_1(\xi_t,t)$ term is found by using the following relationships

$$\begin{aligned}
\frac{\partial A}{\partial \xi_t} &= A\frac{T}{t(T-t)}x\hat{g}_t \\
\frac{\partial B}{\partial \xi_t} &= \int_0^1 A\frac{T}{t(T-t)}x\hat{g}_t\mathrm{d}x = C\frac{T}{t(T-t)}\hat{g}_t
\end{aligned} \tag{4.116}$$

and applying the quotient rule to $\phi_t(x)$

$$\begin{aligned}\Gamma_1(\xi_t, t) = \frac{\partial}{\partial \xi_t}\phi_t(x) = \frac{\partial}{\partial \xi_t}\frac{A}{B} &= \frac{1}{B}\frac{\partial A}{\partial \xi_t} - A\left(\frac{1}{B^2}\right)\frac{\partial B}{\partial \xi_t} \\ &= \phi_t(x)\frac{T}{t(T-t)}\hat{g}_t\left(x - \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t\right]\right) \\ &= \phi_t(x)\frac{T}{t(T-t)}\hat{g}_t\left(x - S(\xi_t, t)\right).\end{aligned} \tag{4.117}$$

The $\Gamma_2(\xi_t, t)$ term is found first by writing

$$\begin{aligned}\frac{\partial A}{\partial t} &= A\frac{\partial}{\partial t}\left(\frac{T}{t(T-t)}\hat{g}_t x(\xi_t - \tfrac{1}{2}\hat{g}_t x)\right) \\ &= A\frac{Tx}{t^2(T-t)^2}\left(\dot{\hat{g}}_t t(T-t)(\xi_t - \hat{g}_t x) - \hat{g}_t(\xi_t - \tfrac{1}{2}\hat{g}_t x)(T-2t)\right)\end{aligned} \tag{4.118}$$

and

$$\begin{aligned}\frac{\partial B}{\partial t} &= \int_0^1 A\frac{\partial}{\partial t}\left(\frac{T}{t(T-t)}\hat{g}_t x(\xi_t - \tfrac{1}{2}\hat{g}_t x)\right)\mathrm{d}x \\ &= \frac{T\dot{\hat{g}}_t}{t(T-t)}\left(\xi_t\int_0^1 xA\mathrm{d}x - \hat{g}_t\int_0^1 x^2 A\mathrm{d}x\right) \\ &\quad - \frac{T(T-2t)\hat{g}_t}{t^2(T-t)^2}\left(\xi_t\int_0^1 xA\mathrm{d}x - \tfrac{1}{2}\hat{g}_t\int_0^1 x^2 A\mathrm{d}x\right);\end{aligned} \tag{4.119}$$

where the dot in $\dot{\hat{g}}_t = \frac{\partial \hat{g}}{\partial t}$ is Newton's dot notation. Notice that the function A has the following properties

$$\int_0^1 xA\mathrm{d}x = B\mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t\right] = S(\xi_t, t) \tag{4.120}$$

and

$$\int_0^1 x^2 A\mathrm{d}x = B\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right]. \tag{4.121}$$

Applying the quotient rule to $\Gamma_2(\xi_t, t)$

$$\begin{aligned}\Gamma_2(\xi_t, t) &= \frac{\partial}{\partial t}\phi_t(x) = \frac{\partial}{\partial t}\frac{A}{B} \\ &= \frac{1}{B}\frac{\partial A}{\partial t} - A\left(\frac{1}{B^2}\right)\frac{\partial B}{\partial t}\end{aligned} \tag{4.122}$$

and using Eq. (4.118) and Eq. (4.121) gives

$$\begin{aligned}\Gamma_2(\xi_t, t) = \phi_t(x)\frac{T}{t^2(T-t)^2}&\Bigg(x\left(\dot{\hat{g}}_t t(T-t)(\xi_t - \hat{g}_t x) - \hat{g}_t(\xi_t - \tfrac{1}{2}\hat{g}_t x)(T-2t)\right) \\ &- \Big(\dot{\hat{g}}_t t(T-t)\left(\xi_t S(\xi_t, t) - \hat{g}_t \mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right]\right) \\ &- \hat{g}_t(T-2t)\left(\xi_t S(\xi_t, t) - \tfrac{1}{2}\hat{g}_t \mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right]\right)\Big)\Bigg).\end{aligned} \tag{4.123}$$

The $\Gamma_3(\xi_t, t)$ term is found by

$$\begin{aligned}\Gamma_3(\xi_t, t) &= \frac{\partial}{\partial \xi_t}\Gamma_1(\xi_t, t) \\ &= \frac{T}{t(T-t)}\hat{g}_t\left(\phi_t(x)\frac{T}{t(T-t)}\hat{g}_t\left(x - \mathbb{E}^{\mathbb{Q}}\left[X_T \mid \xi_t\right]\right)^2 - \phi_t(x)\frac{\partial}{\partial \xi_t}\frac{C}{B}\right)\end{aligned} \tag{4.124}$$

and applying the quotient rule

$$\begin{aligned}\frac{\partial}{\partial \xi_t}\frac{C}{B} &= \frac{1}{B}\frac{\partial}{\partial \xi_t}C - \frac{C}{B^2}\frac{\partial}{\partial \xi_t}B \\ &= \frac{T}{t(T-t)}\hat{g}_t \mathbb{V}^{\mathbb{Q}}\left[X_T \mid \xi_t\right] .\end{aligned} \tag{4.125}$$

Now substituting this into Eq. (4.124), one finds

$$\Gamma_3(\xi_t, t) = \frac{T^2}{t^2(T-t)^2}\hat{g}_t^2\phi_t(x)\left((x - S(\xi_t, t))^2 - \mathbb{V}^{\mathbb{Q}}\left[X_T \mid \xi_t\right]\right). \tag{4.126}$$

Hence, we have found $\Gamma_1(\xi_t,t)$, $\Gamma_2(\xi_t,t)$ and $\Gamma_3(\xi_t,t)$ and one can plug these values into Eq. (4.113) and calculate the three integrals

$$
\begin{aligned}
\int_0^1 x\Gamma_1(\xi_t,t)\mathrm{d}x &= \frac{T}{t(T-t)}\hat{g}_t\mathbb{V}^{\mathbb{Q}}\left[X_T \mid \xi_t\right] \\
\int_0^1 x\Gamma_2(\xi_t,t)\mathrm{d}x &= \frac{T}{t^2(T-t)^2}\Bigg(\dot{\hat{g}}_t t(T-t)(\xi_t\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right] - \hat{g}_t\mathbb{E}^{\mathbb{Q}}\left[X_T^3 \mid \xi_t\right]) \\
&\quad - \hat{g}_t(T-2t)(\xi_t\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right] - \tfrac{1}{2}\hat{g}_t\mathbb{E}^{\mathbb{Q}}\left[X_T^3 \mid \xi_t\right]) \\
&\quad - S\Big(\dot{\hat{g}}_t t(T-t)\left(\xi_t S(\xi_t,t) - \hat{g}_t\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right]\right) \\
&\quad - \hat{g}_t(T-2t)\left(\xi_t S(\xi_t,t) - \tfrac{1}{2}\hat{g}_t\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right]\right)\Big)\Bigg) \\
\int_0^1 x\Gamma_3(\xi_t,t)\mathrm{d}x &= \frac{T^2}{t^2(T-t)^2}\hat{g}_t^2\Bigg(\mathbb{E}^{\mathbb{Q}}\left[X_T^3 \mid \xi_t\right] \\
&\quad - 2S(\xi_t,t)\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right] \\
&\quad + S^3(\xi_t,t) - S(\xi_t,t)\mathbb{V}^{\mathbb{Q}}\left[X_T \mid \xi_t\right]\Bigg).
\end{aligned}
\tag{4.127}
$$

Substituting these three expressions in Eq. (4.127) into Eq. (4.113) we find

$$
\begin{aligned}
\mathrm{d}S_t &= \frac{T}{t(T-t)}\hat{g}_t\mathbb{V}^{\mathbb{Q}}\left[X_T \mid \xi_t\right]\mathrm{d}\xi_t + \\
&\frac{T}{t^2(T-t)^2}\Bigg(\hat{g}_t(T-t)\left(\hat{g}_t - \dot{\hat{g}}_t t\right)\left(\mathbb{E}^{\mathbb{Q}}\left[X_T^3 \mid \xi_t\right] - S(\xi_t,t)\mathbb{E}^{\mathbb{Q}}\left[X_T^2 \mid \xi_t\right]\right) \\
&- \mathbb{V}^{\mathbb{Q}}\left[X_T \mid \xi_t\right]\left(\xi_t(\hat{g}_t(T-2t) - \dot{\hat{g}}_t t(T-t)) + T\hat{g}_t^2 S(\xi_t,t)\right)\Bigg)\mathrm{d}t
\end{aligned}
\tag{4.128}
$$

which is the dynamic price process for a continuous $X_T \in [0,1]$. For a discrete state $X_T \in \{0,1\}$ one finds a simpler form of Eq. (4.128)

$$
\begin{aligned}
\mathrm{d}S_t = \frac{T}{t(T-t)}\hat{g}_t\mathbb{V}^{\mathbb{Q}}\left[X_T \mid \xi_t\right]\Bigg(&\mathrm{d}\xi_t + \frac{1}{t}\left(\hat{g}_t - \dot{\hat{g}}_t t\right) \\
&- \frac{1}{t(T-t)}\left(\xi_t((T-2t) - \frac{\dot{\hat{g}}_t t}{\hat{g}_t}(T-t)) + T\hat{g}_t S(\xi_t,t)\right)\mathrm{d}t\Bigg)
\end{aligned}
\tag{4.129}
$$

where it can be computationally shown for certain definitions of $\hat{g}_t$, such as $\hat{g}_t = t^2$, that

$$\mathrm{d}\tilde{B}_t = \mathrm{d}\xi_t - \frac{1}{t(T-t)}\left(\xi_t\left((T-2t) - \frac{\dot{\hat{g}}_t}{\hat{g}_t}t(T-t)\right) + T\hat{g}_t S(\xi_t, t) + \frac{1}{t}(\hat{g}_t - \dot{\hat{g}}_t t)\right)\mathrm{d}t \tag{4.130}$$

is a standard Brownian motion. Finally one arrives at the generalised dynamic price process

$$\mathrm{d}S_t = \frac{T}{t(T-t)}\hat{g}_t\left(1-\phi_{1t}\right)\phi_{1t}\mathrm{d}\tilde{B}_t \tag{4.131}$$

which is applicable if and only if $\lim_{t\to T}\frac{g(t)}{\mathbb{V}[\beta_{tT}]} \to \infty$. Notice that Eq. (4.131) yields the BHM model by setting $\hat{g}_t = \sigma t$ which reproduces the same price dynamics as in Eq. (4.62). Now that Eq. (4.131) has been derived one can now use this form to fit to real data which is performed in Sec. 4.7.2.

4.7 Fitting Racing Data to An Information Process

Using the theory reviewed and developed in Sec. 4.2.3 the BHM model is fitted to the *last price matched* signal, see Sec. 2.2, for the winning horses. We first fit the linear model, as defined by the original BHM model as in Eq. (4.11), in Sec. 4.7.1. The next fit that is performed is with the non-linear information model that was developed in Sec. 4.6, derived as Eq. (4.131), and this is described in Sec. 4.7.2.

4.7.1 Fitting a Linear Information Rate Model

This section explores the application of a fitting procedure to the linear information rate BHM-model, Eq. (4.11), which is based on

$$\xi_t = \hat{\sigma}X_T t + \beta_{tT} \tag{4.132}$$

where $\hat{\sigma}$ is the positive constant information rate parameter that will be fitted and all other notation is the same as in Sec. 4.2. The fitting of $\hat{\sigma}$ is applied to the winning last price matched signals ($X_T = 1$) which are described in the Table 2.1 and Sec. 2.2. The winning last price matched signals are filtered and sorted by the course distances, splitting the data into sub-samples. There tends to be a period of lower fluctuation in the last price matched winning signal at the start of the race compared to when the volatility cluster is observed, see Fig. 2.4 middle left. A volatility cluster is defined as tendency of large price movements to clump together, this is illustrated in Fig. 4.19 using the return series from the S&P500.

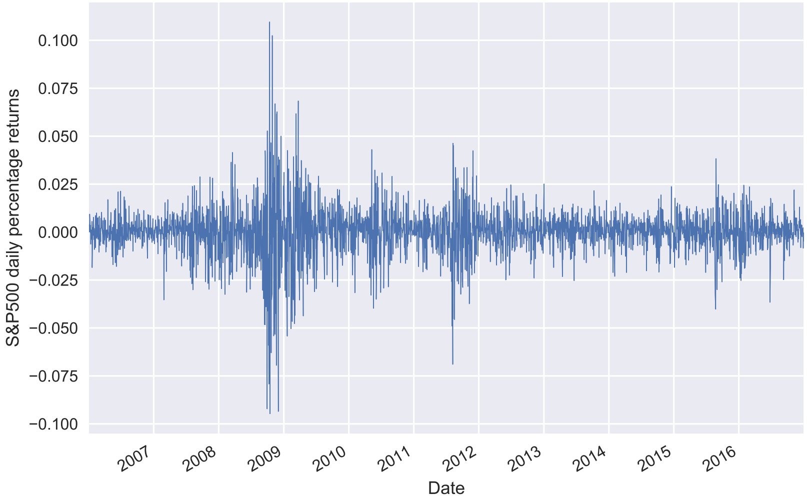


Figure 4.19: This is the percentage return series for the S&P500 starting from 01/01/06 and ending 31/12/16. We see there are parts of the signal that oscillate with large swings, as compared to the majority of the series, and these are clumped together. This is known as a volatility cluster and the most prevalent one is found round the period 2008-2009.

The volatility cluster found with the gambling data is assumed to be the point at which the information process is initiated, such that the time coordinate in the model is defined to start at this point. To find the time when the volatility cluster starts to emerge we use an exponential weighted moving average model (EWMA) [18,81] which is calibrated on the increments of the last price matched: this running volatility measure is then compared to the signal's standard deviation. If the EWMA measure of volatility exceeds the standard deviation then this is the point at which the volatility cluster starts and the point $t = 0$ in the calibration of our BHM model fit. The sub-samples determined by the race distance are represented by 100 randomly picked winning signals which are averaged by binning and averaging the data points in those bins; these bins are equally spaced in the interval $t \in [0,1]$. These averaged signals are then shifted up vertically so that the final value of price is 1 and then rotated about the end data point, which is the point ($t = 1$, Last Price Match = 1), ensuring that all the signals start at the same a priori probability (denoted as p_1 in the BHM model Eq. (4.51)). The fit is performed on each of the average price signals that are designated by the course distance. We define the information rate parameter space as the interval $\hat{\sigma} = [0.5, 2]$ which is split into evenly spaced values of step size 0.01. We define there to be $N_\sigma = 151$ values of the $\hat{\sigma}$ parameter in the search space. We index the interval $\hat{\sigma} = 0.5, 0.51, 0.52, \ldots, 2$ as $\hat{\sigma}_l = 0.5 + l(0.01)$ where $l = 0, 1, 2, \ldots 150$. For each value of $\hat{\sigma}_l$ one generates 1.0×10^4 information processes and thus 1.0×10^4 price signals using the BHM pricing kernel in Eq. (4.19). These price processes are denoted as $\{S(t, \hat{\sigma}_l)\}_{u=1,2,\ldots,1\times10^4}$ for each $l = 0, 1, 2, \ldots 150$ and then they are averaged over the 1.0×10^4 samples, denoted $\bar{S}_t^{(l)} = \langle S(t, \hat{\sigma}_l) \rangle$. The average empirical price is $J_t^{(j)} = \left\langle LPM_t^{(j)} \right\rangle$ where the index $j = 1, 2, \ldots N_d$ corresponds to the sub-samples filtered on course distance and $\langle . \rangle$ is the binned average over the 100 signals. The mean square error (MSE) on the set of parameters $\hat{\sigma}_l$ is defined as

$$\{MSE\left(\hat{\sigma}_l\right)\}_{j=1,2,\ldots,N_d} = \frac{1}{T} \sum_{\tau=1}^{T} \left(J_\tau^{(j)} - \bar{S}_\tau^{(l)} \right)^2 \tag{4.133}$$

where $l = 0, 1, \ldots, 150$, $\tau = 1, 2, \ldots, T$ is indexed time t within the averaged binned price $J_t^{(j)}$ and the average BHM price is $\bar{S}_t^{(l)}$. Using the MSE as the fitting measure, one can find the $\hat{\sigma}_l$ for each race distance $j = 1, 2, \ldots, N_d$ that gives the smallest value of MSE. This procedure is defined mathematically as the minimum mean square error (MMSE)

$$\hat{\sigma}_j = \left\{ \arg\min_{\sigma_l} \left(MSE\left(\sigma_l\right)\right) \right\}_{j=1,2,\ldots,N_d} \tag{4.134}$$

hence we estimate the $\hat{\sigma}_j$ for each race distance j. This process is repeated 100 times and $\hat{\sigma}_j$ for each $j = 1, 2, \ldots, N_d$ is shown in Fig. (4.20). On a log-log scale one observes a linear relationship between $\log\left(\hat{\sigma}_j\right)$ and the logarithm of race course distance. A linear relationship on a log-log scale is a power-law where the gradient of the linear line is the exponent of the power law which is estimated as $\hat{m} = 0.3993 \pm 0.0413$. This positive value of $\hat{m}$ indicates that races of a smaller distance are dominated more by noise than the longer races.

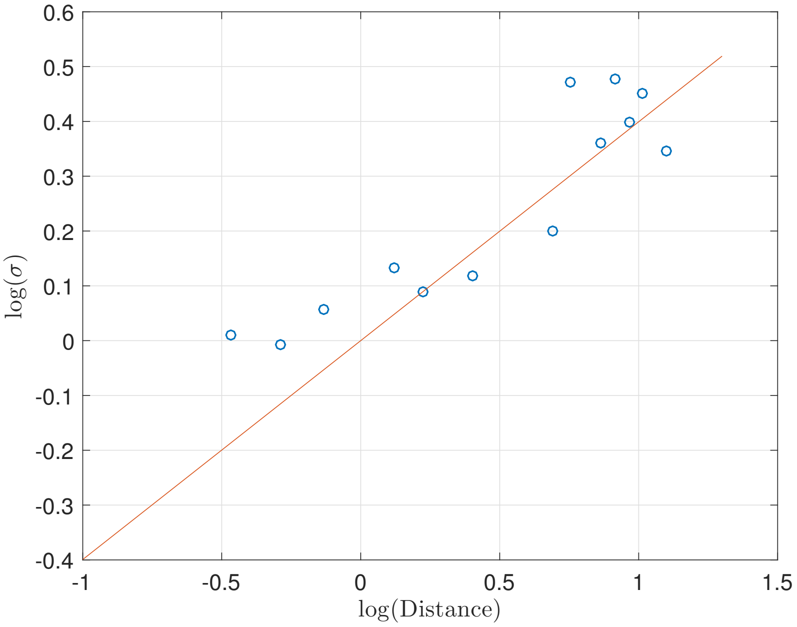


Figure 4.20: A log-log plot where the blue circles are the estimated values of $\hat{\sigma}_j$ for each of the distances $j = 1, 2, \ldots, N_d$. The red line is a linear model with intercept, the gradient is estimated to be $\hat{m} = 0.3993 \pm 0.0413$. This estimation of $\hat{m}$ has a p-value of 5.18×10^{-7} and $R^2 = 0.8164$ both indicating that this linear model is statistically significant.

An issue when calculating the variance process Eq. (4.59) with the fitted linear information rate $\hat{\sigma}$ is demonstrated in Fig. 4.21. One observes from Fig. 4.21 that the model underestimates the variance after the time $t \approx 0.2$. The volatility profiles in Fig. 4.21 of the model (blue line) and the data (red line) do have a similar structure but we can improve the fit by employing a non-linear information process in Sec. 4.7.2.

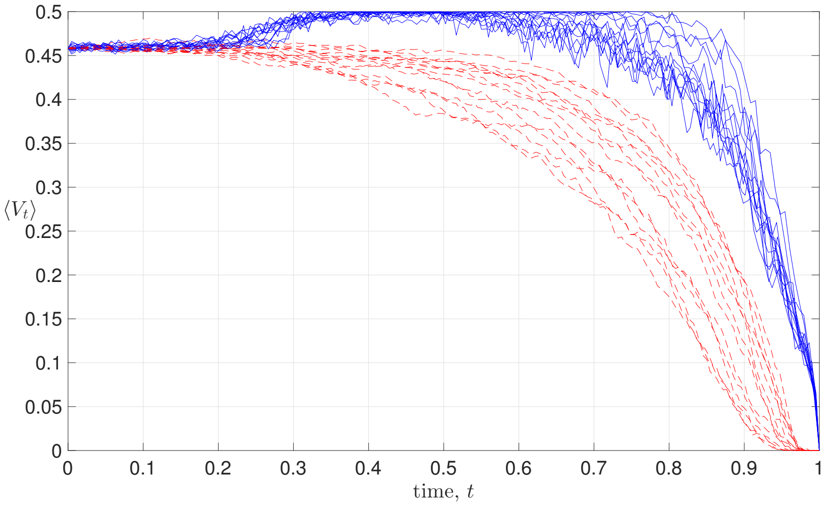


Figure 4.21: The solid blue lines are the average variance calculated with the real price data. The average variance is calculated by Eq. (4.59) and averaged over 10 samples. The dashed red line is the average variance for the simulated BHM model with $\hat{\sigma}_j$ given by the blue circles in Fig. 4.20. Each of the lines in the two stack (red and blue) from top to bottom are increasing in race distance, and are estimates for the variance.

4.7.2 Fitting a Non-Linear Information Rate Model

This section explores the implementation of the non-linear information process defined in Eq. (4.107). This is of the form

$$\xi_t = \hat{g}(t) X_T + \beta_{tT} \tag{4.135}$$

where $\hat{g}(t)$ is a function that will be estimated. The price of such an asset with a binary terminal payoff is found in Sec. 4.2.3 to be

$$S_t = \left(1 + \tfrac{p_1}{p_0} \exp\left(-\tfrac{T}{t(T-t)}\hat{g}(t)\left(\beta_{tT} + \tfrac{1}{2}\hat{g}(t)\right)\right)\right)^{-1} \tag{4.136}$$

where $0 \le t < T$, $0 < S_t < 1$, p_0 and p_1 are the a priori probabilities as defined in Eq. (4.51). The price Eq. (4.136) is inverted and a sample average taken which gives

$$\hat{g}(t)(\langle\beta_{tT}\rangle_N + \tfrac{1}{2}\hat{g}(t)) - \langle h(t, S_t)\rangle_N = 0 \tag{4.137}$$

where function $h(t, S_t)$ is defined as

$$h(t, S_t) \triangleq -\frac{t(T-t)}{T} \ln\left(\frac{p_1}{p_0}\frac{(1-S_t)}{S_t}\right) \tag{4.138}$$

such that $h(t = 0, S_0 = p_1) = 0$ and $h(t = T, S_T = 1) = 0$. As stated in Sec. 4.1.3 the

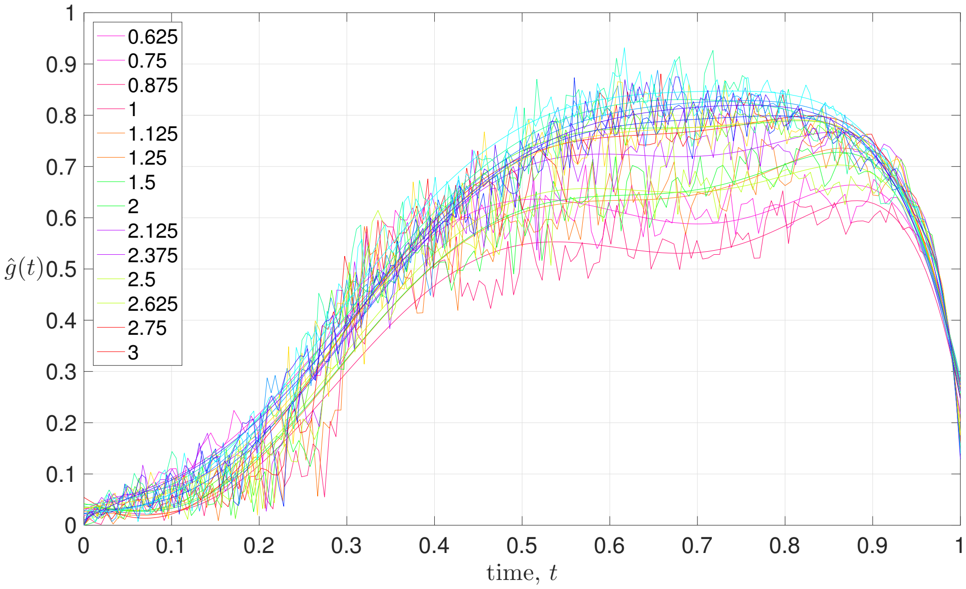


Figure 4.22: The implementation of the non-linear function fitting procedure as defined by Eq. (4.139). The noisy solid lines that fluctuate are fits obtained using Eq. (4.139) and the smooth solid lines are a degree 6 polynomial that has been calibrated to fit Eq. (4.139). Each colour denotes price data that has been sub-categorised by the distance shown (in miles) in the legend.

Brownian bridge has the following property $\langle \beta_{tT} \rangle_N = 0$ and Eq. (4.137) becomes

$$\hat{g}(t) = \sqrt{2 \langle h(t, S_t) \rangle_N} \tag{4.139}$$

where $\langle . \rangle_N$ is the average over the sample of size N. We use the same empirical price data as in the linear fit model, Sec. 4.7.1, which is denoted as $J_t^{(j)} = \left\langle LPM_t^{(j)} \right\rangle$ and fitted with Eq. (4.139) where $\bar{S}_t^{(j)} = \langle S(t, \hat{g}_j(t)) \rangle$ and $j = 1, 2, \ldots N_d$ as before. The results are displayed in Fig. 4.22. It is clear from this plot that the information rate function is not linear. We estimate the corresponding of variances in Fig. 4.23 using the same method as explained in the previous section Sec. 4.7.1.

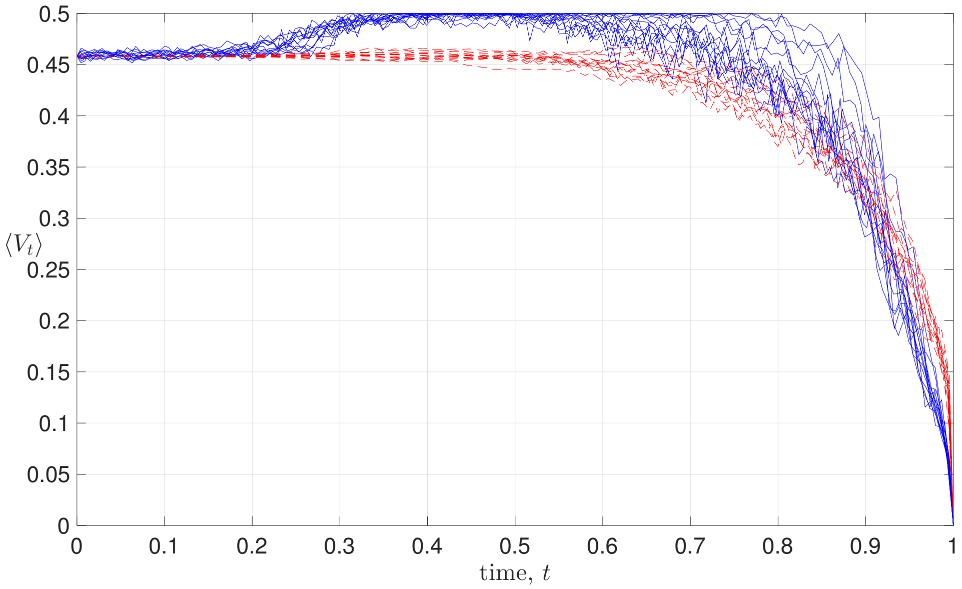


Figure 4.23: The solid blue lines are the average variance calculated with the real price data. The average variance is calculated by Eq. (4.59) and averaged over 10 samples. The dashed red line is the average variance for the simulated non-linear BHM model with $\hat{g}(t)$ given by the smooth lines in Fig. 4.22.

4.8 Summary and Discussion

In this chapter we discussed and reviewed the main body of literature where the BHM framework was developed [86–110]. This BHM framework is discussed in detail in Sec. 4.1 where the information process is mathematically defined as the union of the true information rate and Brownian bridge price. We then show in Sec. 4.2 that by applying the BHM framework to model filtrations one can price financial instruments. This pricing schema is then used to derive a dynamic form of price evolution, see Eq. (4.30), where we notice that a Brownian process drives the noise but on the other hand generates the market information. We also shown the necessary limiting condition for the conditional density function in Eq. (4.70) which if broken means the price will not converge to correct values.

In Sec. 4.3 we developed the BHM framework such that a synthetic marketplace can be achieved. We first created a mechanism for trading to take place and proposed how the prices of the trading agents update after trading. The agents are categorised as either market making, informed traders and noise traders where each type of agent has a different trading mechanism. To control the number of trades made by the market makers we defined the inventory control mechanism in Sec. 4.91 and this is applied so we can observe the effect on the market price arising from the informed traders and noise traders. The results for the trading model are presented in Sec. 4.4 where we present the distributions for the market returns and first passage times for the following configurations $(2, 0, 0)$, $(2, 1, 0)$ and $(2, 0, 1)$. For the $(2, 0, 0)$ configuration we change parameters such as the spread which when increased reduces the frequency of trading and increases the standard deviation of the returns. For the configuration $(2, 1, 0)$ we change the spread of the informed trader and apply inventory control and find the tentative result that a bimodal distribution is observed which is something that needs further examination. For the configuration $(2, 0, 1)$ we change the threshold parameter denoted as γ and defined in Eq. (4.90), we find that increasing γ decreases the frequency of trading.

The Sec. 4.5 explores the introduction of multiple information processes with the example of horse racing markets. We find that making the information processes independent ensure that the limiting property Eq. (4.70) holds true. This is a somewhat frustrating condition because it seems intuitive that information processes that are correlated would give a richer framework for collections of agents pricing the same asset.

The Sec. 4.6 discusses a way of introducing a non-linear rate function instead of a linear function as proposed in the BHM model. We defined this as a general function that obeys the condition Eq. (4.70). We apply Itô's Lemma to this non-linear version of the information process and derive the dynamical price process Eq. (4.131). We find that this process has a similar structure to the BHM price model but now one can try to fit this non-linear information rate function to see if it is linear or not.

The final section, Sec. 4.7 took the original BHM theory and successfully applied it to the winning horse signals from the data set. We found that when fitting to the average last price matched using the original BHM model the information rate parameter followed a power law with a positive exponent against course distance, see Fig. 4.20. This gives the interpretation that the odds in short races are governed more by noise than is found in longer races. We took the non-linear BHM model derived in Sec. 4.6 and fitted it to the same data set as the linear BHM. This estimation of the information rate function does not produce a linear function but one that can be regarded as either piecewise linear or a polynomial of order 6.

Chapter 5

Optimal Trading Strategies – a Time Series Approach

The last couple of decades have seen many physicists becoming interested in the question of portfolio optimisation [133–144]. Key issues addressed in these studies concern the effects that sampling noise is likely to have on the measurement of correlations or covariances in large portfolios, the way in which such sampling noise is going to affect the solution of a subsequent mean-variance portfolio optimisation problem, and the design of methods to mitigate against adverse effects of such sampling noise.

The bedrock of most of these studies is the theory of random sample covariance matrices [145]. Their spectral theory was pioneered by Marčenko and Pastur [146] in the 1960's. It has indeed been observed that, apart from a number of *large* eigenvalues, the bulk of the spectrum of sample covariance matrices of asset returns in various markets is very close to the form predicted by Marčenko and Pastur for sample covariance matrices of i.i.d. random data; see e.g. [133, 134]. This type of comparison between market data and a null-model defined by random data could then be used to devise theory-guided ways of distinguishing between information and noise in market data, and thereby to devise methods to clean covariance matrices of asset returns for the purpose of their subsequent use in portfolio optimisation, with the effect of improving risk-return characteristics [133, 135–144, 147].

The present study was triggered by the fact that the spectral theory of sample *auto-covariance* matrices, the analogue of [146] in the time domain has recently become available [148]. This leads us to revisit the analogue of Markowitz mean-variance optimisation in the time domain [6], which in its simplest incarnation allows to find an optimal trading strategy for a single traded asset over a finite (discrete) time horizon. We investigate this setup for a range of synthetic processes, taken to be either second order stationary, or to exhibit second order stationary increments, and we systematically study the effects of sampling noise on optimal strategies and on risk-return characteristics. Finally, we apply our framework to daily returns of the S&P500 index, and we explore how results obtained for spectra of sample auto-covariance matrices obtained in [148] could then be used as a guide to clean sample auto-covariance matrices in a spirit analogous to that used for sample-covariance matrices in the context of portfolio optimisation.

We note at the outset that we regard this as an exploratory study, and that we ignore economic factors such as discounting and agents' asymmetric perceptions of gains and losses in the present study. Expecting that the primary area of application of our techniques would be in the high-frequency domain, as return auto-correlations will be most prominent at short times. We note, however, that much of our analysis is about the effects of sampling noise on optimal trading strategies, which is relevant at *all* time scales, and thus also for weakly correlated data.

5.1 Time Series Review

Times series can be exercised in a multitude of areas of applied mathematics and can be a powerful tool in real time statistical analysis. Using time series to split a signal into a component sum of deterministic trends, seasonal trends and random fluctuations (i.e. Gaussian white noise). Trends are removed from a data set by procedures such as *differencing*, *de-trending* and *Hodrick-Prescott filter*. These normalising methods ensures that no long term trends are mistaken for seasonal movements also known as "*seasonal adjustment*". A beneficial property of time series modelling is forecasting; the technique filters out random components of the signal and find hidden deterministic trends that can be applied to the future to make predictions. A reason for adopting time series for this study is the synthetic generation of stochastic signals, which will be assumed to mimic the fluctuation of an assets' price in time (such as an index or component of an index). Times series models can be calibrated by maximum-likelihood or machine learning techniques [149], with real data which in turn can be used as a means to simulate the fluctuations of real data. This calibration may have some predictive power for some time scales, but these have been highly criticised in some of the economics literature [150]. The signal is simulated by an underlying process such as an auto-regressive process and in turn generates synthetic data for analysis. Other examples of random signals could be tidal levels [76, 151], import/export quantities [152, 153], and pairs trading [47–52] and Sec. 3.3. Adopting time series methodologies gives a means of simulating synthetic processes and stress test the models developed. Depending on subjective opinion of the model builder with information criteria [154] and test-statistics, one can build a model such as ARIMA(p, i, q) or GARCH(p, q) to model price or return fluctuations.

5.1.1 Mathematical Framework

This section outlines the fundamental mathematical definitions of time series analysis used in our model that can be also found in the following textbooks [18, 81, 155–157]. A time series can be thought of as a sequence of random variables that have been sampled or measured in time, known as a time index. The time index is a member of a time index set $\mathbb{T}$, for example $\mathbb{T} \subset \mathbb{Z}$. The time index set is by no means a strict definition of $\mathbb{T}$, as one can index with the continuous real line $\mathbb{R}$ (see [68]). This chapter will only be concerned with the discrete time index case that is $t \in \{1, 2, ..., T\}$. The random variables of a time series are defined in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the sample space, $\mathcal{F}$ is the σ-field and $\mathbb{P}$ is the physical probability measure. We have now denoted enough variables to define a stochastic signal and a time series as

$$\begin{aligned} &\{X_t(\omega) \mid \omega \in \Omega, t \in \mathbb{T}\} \text{ - Stochastic Signal/Process} \\ &\{x_t \mid t \in \mathbb{T}\} \text{ - Sequence of observations of } X_t(\omega) \text{ (filtration).} \end{aligned} \tag{5.1}$$

This sequence of random variables is assumed to be distributed in some manner and will be simple and well behaved. If the event is fixed $\omega \in \Omega$ and the time index set is $t = \{1, 2, ..., T\}$, then the discrete sample path of the underlying stochastic signal is $\{X_1, X_2, ..., X_T\}$. Each realisation of $\{X_t(i)\}_{i=1}^{M}$ gives a single value and one can repeat the operation of measurement which supplies M samples denoted as

$$\{X_1(1), X_2(1), ..., X_T(1)\}, \{X_1(2), X_2(2), ..., X_T(2)\}, ..., \{X_1(M), ..., X_T(M)\}.$$

Estimating the expectation of $\{X_t(i)\}$, from this sample space, is found by taking the ensemble average for all the realisation at time t. In reality one only observes a particular sample path for the underlying process, so it is insurmountable to calculate the sample mean for all i. A more prudent way of estimating the expectation of a stationary process $\{X_t\}$ is the time average

$$\langle X_t \rangle = \tfrac{1}{M} \sum_{t=1}^{M} x_t \ . \tag{5.2}$$

The time average of a stochastic signal is only feasible if the probability distribution of the underlying process $\{X_t\}$ is *time-shift* invariant also known as signal *stationarity*.

5.1.2 Stationary Signals

The stationarity property is at the heart of time series and allows one to exercise statistical analysis on a stochastic time signal $\{X_t\}$. Non-formally, a stationary signal is one that has similar or unchanged statistical properties under time translation (time-shift invariant). Formally, a stationary signal $\{X_t\}$ is one that has an independent joint cumulative distribution function (CDF) under a time-shift, defined in Eq. (5.3).

5.1.3 Strict Stationarity

Consider a stochastic signal $\{X_t \mid t \in \mathbb{T}\}$ which for any subset of the time index set $\mathbb{T} = \{t_{-n}, t_{-n+1}, \ldots, t_{-2}, t_{-1}, t_0, t_1, t_2, \ldots t_{n-1}, t_n\}$. For any lag $h \in \mathbb{T}$ such that $t_i + h \in \mathbb{T}$ where $\{i = -n, -n+1, \ldots, -2, -1, 0, 1, 2, \ldots, n-1, n\}$ and n being a positive integer; if the CDF of the process $\{X_t \mid t \in \mathbb{T}\}$ is invariant under the following time-shift

$$F_X(x_{t_{-n}}, ..., x_{t_{-1}}, x_{t_0}, x_{t_1}, ..., x_{t_n}) = F_X(x_{t_{h-n}}, ..., x_{t_{h-1}}, x_{t_h}, x_{t_{h+1}}, ..., , x_{t_{h+n}}) \tag{5.3}$$

then $\{X_t \mid t \in \mathbb{T}\}$ is said to be *strictly stationary*. This implies the distribution F_X is independent of the time-shift and the signal $\{X_t\}$ is strictly-stationary. One can express the definition of strict-stationarity explicitly in terms of time dependent expectation and auto-covariance:

(i) $\mathbb{E}[X_t] = \mu_X(t)$.

(ii) $\mathbb{C}[X_t, X_{t'}] = \mathbb{E}[(X_t - \mu_X(t))(X_{t'} - \mu_X(t'))] = \gamma_X(t, t'); \quad \forall t, t' \in \mathbb{T}$.

Here it has been assumed that the signal is defined in terms of the $\mathcal{L}_2$-norm[*] such that the process $\{X_t \in \mathcal{L}_2 \mid t \in \mathbb{T}\}$ has the following finite expectation $\mathbb{E}[|X_t|^2] < \infty$. This assumption ensures that the two properties, (i) and (ii), from the above box are both finite and exist. The function $\gamma_X(t, t')$ is known as the auto-covariance function and $\gamma_X(t, t) = \mathbb{V}(X_t)$.

5.1.4 Weak Stationarity

The strict-stationarity property is usually too strong, rigid and impractical of a property to test on a real time-series, as an empirical distribution is never completely realised. In the analysis of empirical data, one would find it extremely difficult, nigh impossible, to ensure strict-stationarity is maintained throughout the signal. A more pragmatic and practical approach is to adopt a weaker form of strict-stationarity, a statistical approach based on the signals first and second

[*] The general p-norm is defined as $X_t \in \mathcal{L}_p \Leftrightarrow \mathbb{E}[|X_t|^p] < \infty$.

moments. If one assumes that the mean and covariance function of the signal $\{X_t \in \mathcal{L}_2 \,|\, t \in 1, 2, \ldots, M\}$ hold statistically in the following way:

$$\boxed{\begin{aligned} &\text{(i)}\;\; \mu_X(t) = \mu_X(t') = \hat{\mu}_X, \text{ where } t' = t + h \;\; \forall h \in \mathbb{T} \\ &\text{(ii)}\;\; \mathbb{C}[X_t, X_{t'}] = \gamma_X(t, t') = \hat{\gamma}_X(h) \end{aligned}} \tag{5.4}$$

then one has a weak stationarity. The hat indicates that a statistical estimation of the moment is being performed, see Eq. (5.7). The definitions of a weak stationary signal shows that the mean and covariance are statistically time-shift invariant which can be tested with a Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) test [75, 81, 158].

Consider a stochastic signal $\{X_t\}$ which is weak stationary, one can define the auto-correlation function as

$$\hat{\rho}_X(h) = \frac{\hat{\gamma}_X(h)}{\hat{\gamma}_X(0)} \tag{5.5}$$

where $\hat{\rho}_X(h) \in [-1, 1]$ is bounded.

The fact that the empirical auto-correlation and auto-covariance function are even symmetric functions, implies one can define the Toeplitz form as the following matrix

$$\{\boldsymbol{\Sigma}\}_{ij} = \gamma_X\left(|i - j|\right) \tag{5.6}$$

where the auto-covariance matrix is the true auto-covariance matrix if the number of sample $M \to \infty$.

5.1.5 The Empirical Estimation

The empirical mean and variance are defined respectively for M observation as

$$\begin{aligned} \hat{\mu}_X &\triangleq \frac{1}{M}\sum_{t=1}^{M} x_t \\ \hat{\sigma}_X^2 \triangleq \hat{\gamma}_X(0) &= \frac{1}{M}\sum_{t=0}^{M-1}(x_t - \hat{\mu}_X)^2 \end{aligned} \tag{5.7}$$

where the hat represents one is estimating the parameters with a finite sample size $M < \infty$. The empirical auto-covariance matrix is defined as

$$\hat{\Sigma}_{tt'} = \frac{1}{M}\sum_{k=0}^{M-1}(x_{k+t} - \langle x_{k+t}\rangle)((x_{k+t'} - \langle x_{k+t'}\rangle), \quad \text{for } 1 \le t, t' \le T \tag{5.8}$$

where this definition is taken from [148] and we have that the sampling error disappears in the following limit $\lim\limits_{M\to\infty}\{\hat{\mathbf{\Sigma}}\} = \mathbf{\Sigma}$.

5.1.6 White Noise Process

Consider a stochastic signal $\{X_t \mid t = 1, 2, 3, \dots\}$ and $X_0 = 0$; the signal for each time is independent and identically distributed ($i.i.d$) as a Gaussian and has a second order moment $\mathbb{E}[|X_t|^2] = \sigma^2 < \infty$. This Gaussian distribution is centred about the value 0 and has the first order moment $\mathbb{E}[X_t] = 0 \ \forall t$, therefore the weak stationarity conditions holds. One writes this process as $\{X_t\} \sim \mathcal{N}(0, \sigma^2)$ and because this process is $i.i.d$ the generated sequence will have zero cross-correlation for each time t. The auto-covariance function of the white noise process is

$$\gamma_X(h) = \sigma^2 \mathbb{1}_{\{h=0\}} \tag{5.9}$$

where the indicator $\mathbb{1} = 1$ when the lag $h = 0$ and $\mathbb{1} = 0$ otherwise. The true auto-covariance matrix is

$$\mathbf{\Sigma} = \sigma^2 \mathbb{I} \tag{5.10}$$

where $\mathbb{I}$ is the identity matrix with the size $(h \times h)$.

5.1.7 Auto-Regressive AR(p) Process

This process is defined as

$$X_t = a_0 + a_1 X_{t-1} + a_2 X_{t-2} + ... + a_p X_{t-p} + \xi_t; \quad t \in \mathbb{T} \tag{5.11}$$

where $\{\xi_t\} \sim \mathcal{N}(0, \sigma^2)$ $\forall t$ and $a_0, a_1, a_2, ..., a_p$ are the fixed parameters of the model with p being the number of degrees-of-freedom. Using the lag operator as in Eq. (3.27) one can write the AR(p) process as

$$X_t = a_0 + \sum_{k=1}^{p} a_k L^k X_t + \xi_t \tag{5.12}$$

expanding and rearranging we find

$$\left(1 - a_1 L - a_2 L^2 - \cdots - a_p L^p\right) X_t = \phi\left(L\right) X_t = a_0 + \xi_t \tag{5.13}$$

hence,

$$X_t = \phi\left(L\right)^{-1}\left(a_0 + \xi_t\right) \ , \tag{5.14}$$

where the function $\phi\left(L\right)$ is known as the auto-regressive polynomial. Taking the expectation of Eq. (5.13) gives

$$\mathbb{E}\left[X_t\right] = \phi\left(L\right)^{-1}\left(a_0 + \mathbb{E}\left[\xi_t\right]\right) = \phi\left(L\right)^{-1} a_0 \tag{5.15}$$

to ensure the expectation does not explode to infinite, we must have $\phi\left(L\right) \neq 0$. Noticing that the auto-regressive polynomial can rewritten by defining the polynomial variable as $L^k X_t = y^{p-k}$ we arrive at the following

$$y^p - a_1 y^{p-1} - a_2 y^{p-2} - \cdots - a_{p-1} y - a_p \ . \tag{5.16}$$

Hence, for $\phi\left(L\right) \neq 0$ the root of this polynomial must lie within the unit circle if the AR(p) process is to be weakly stationary. This result comes from the fundamental theorem of algebra

$$\phi(y) = (y_1 - y)(y_2 - y) \ldots (y_p - y) \ , \tag{5.17}$$

where the roots are denoted as $y_1, y_2, \ldots, y_p$ and if $|y_i| < 1\ \forall i$ one has weak stationarity. The variance of the AR(p) process is found in a similar way and one finds

$$\mathbb{V}[X_t] = \left(\phi(L)^{-1}\right)^2 \mathbb{V}[\xi_t] = \left(\phi(L)^{-1}\right)^2 \sigma^2 \tag{5.18}$$

giving the same condition that weak stationarity requires that $\phi(L) \neq 0$.

5.1.8 Auto-Regressive AR(1)

The auto-regressive AR(1) process are the simplest of auto-regressive models and defined from Eq. (5.11) as

$$X_t = a_0 + a_1 X_{t-1} + \xi_t; \quad t \in \mathbb{T} \tag{5.19}$$

where a_0 and $|a_1| < 1$ are constants. If one applies the principle of weak stationarity then the expectation of Eq. (5.19) gives

$$\begin{aligned} \mathbb{E}[X_t] &= a_0 + a_1 \mathbb{E}[X_{t-1}] + \mathbb{E}[\xi_t] \\ \mu_X &= a_0 + a_1 \mu_X \Rightarrow \mu_X = \frac{a_0}{1 - a_1}\ , \end{aligned} \tag{5.20}$$

where we define $\mu_X = \mathbb{E}[X_t]$. The Gaussian noise term $\{\xi_t\}$ is also uncorrelated with $\{X_t\}$ for each $t > h$ and if $a_0 = 0$ and the expectation is $\mu_X = 0$. With repetitive substitution of the AR(1) series one finds an iterative expression for X_t as a sum of the parameters and white noise

$$\begin{aligned} X_t &= (a_0 + \xi_t) + a_1(a_0 + a_1 X_{t-2} + \xi_{t-1}) \\ &= (a_0 + \xi_t) + a_1(a_0 + \xi_{t-1}) + a_1^2(a_0 + a_1 X_{t-3} + \xi_{t-2}) \\ &= a_0 + \xi_t) + a_1(a_0 + \xi_{t-1}) + a_1^2(a_0 + \xi_{t-2}) + \ldots \end{aligned} \tag{5.21}$$

$$\Rightarrow X_t = \sum_{i=0}^{\infty} a_1^i (a_0 + \xi_{t-i}). \tag{5.22}$$

If $|a_i| < 1$ one can evoke the geometric sum $\sum_{i=0}^{\infty} a_1^i a_0 = \frac{a_0}{1-a_1} = \mu_X$ and the variance of the AR(1) process is

$$\begin{aligned} \mathbb{V}[X_t] &= \mathbb{V}\left[\sum_{i=0}^{\infty} a_1^i \xi_{t-i}\right] \\ &= \sum_{i=0}^{\infty} \left(a_1^2\right)^i \mathbb{V}[\xi_{t-i}] \end{aligned} \tag{5.23}$$

$$\Rightarrow \mathbb{V}[X_t] = \frac{\sigma^2}{1 - a_1^2} = \sigma_X^2 \tag{5.24}$$

and the auto-covariance function is found as

$$\begin{aligned} \gamma_X(h) &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} a_1^i \xi_{t-i}\right)\left(\sum_{i=0}^{\infty} a_1^i \xi_{t-h-i}\right)\right] \\ &= a_1^h \sum_{i=0}^{\infty} a_1^{2i} \mathbb{E}\left[\xi_{t-h-i}^2\right] \end{aligned} \tag{5.25}$$

$$\Rightarrow \gamma_X(h) = \frac{a_1^h \sigma^2}{1 - a_1^2} = a_1^h \gamma_X(0)\ . \tag{5.26}$$

The AR(1) process, as defined by Sec. 5.19, will be used throughout this chapter as means to generate synthetic data to test the model.

5.2 Portfolio Optimisation: The Markowitz Set-Up

In the simplest version of mean-variance portfolio optimisation one considers a set of N tradable assets $i = 1, \ldots, N$. It is usually assumed that these do not include complex financial instruments such as derivatives, options and futures. An investor can take positions on these assets. We will use π_i to denote the position on asset i, using the convention that $\pi_i > 0$ represents a long position (buying the asset), whereas $\pi_i < 0$ represents a short position (selling asset). With r_i denoting the (random) return on the i^{th} asset, the return on the entire portfolio with positions $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)'$ is given by

$$R(\boldsymbol{\pi}) = \sum_{i=1}^{N} \pi_i r_i = \boldsymbol{\pi}' \boldsymbol{r} \ , \tag{5.27}$$

where $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)'$ is used to denote the vector of random returns and the prime indicates a transpose. The optimal portfolio according to Markowitz is the one that minimises the *variance* of the portfolio return,

$$\mathbb{V}[R(\boldsymbol{\pi})] = \sum_{i,j=1}^{N} \pi_i \pi_j \left\langle (r_i - \mu_i)(r_j - \mu_j) \right\rangle = \sum_{i,j=1}^{N} \pi_i \pi_j \Sigma_{ij} \ , \tag{5.28}$$

subject to the constraint of a given expected portfolio return μ_P

$$\mu_P \equiv \langle R(\boldsymbol{\pi}) \rangle = \sum_{i=1}^{N} \pi_i \langle r_i \rangle = \sum_{i=1}^{N} \pi_i \mu_i \ . \tag{5.29}$$

In Eq. (5.28), $\boldsymbol{\Sigma} = (\Sigma_{ij})$ is the covariance matrix of asset returns. To put a scale to the problem, one usually imposes the normalisation constraint

$$\boldsymbol{\pi}' \mathbf{1} \equiv \sum_{i=1}^{N} \pi_i = 1 \ . \tag{5.30}$$

Here $\mathbf{1} = (1, 1, \ldots, 1)'$ denotes the N dimensional vector with all components equal to 1. The minimisation problem is solved using the method of Lagrange multipliers to take the constraint of expected return and normalisation into account. One is looking for the stationary point for the following Lagrange function

$$\mathcal{L} = \frac{1}{2} \boldsymbol{\pi}' \boldsymbol{\Sigma} \boldsymbol{\pi} - \lambda_1 (\boldsymbol{\pi}' \mathbf{1} - 1) - \lambda_2 (\boldsymbol{\pi}' \boldsymbol{\mu} - \mu_P) \tag{5.31}$$

with respect to variations of the π_i, λ_1 and λ_2. Elementary linear algebra then entails that the optimal portfolio $\boldsymbol{\pi}^*$ takes the form

$$\boldsymbol{\pi}^* = \lambda_1 \Sigma^{-1} \mathbf{1} + \lambda_2 \Sigma^{-1} \boldsymbol{\mu} \ , \tag{5.32}$$

with actual values of the Lagrange parameters λ_1 and λ_2 determined by the constraints.

5.2.1 Translation into the Time-Domain

The Markowitz portfolio optimisation problem allows a relatively straightforward translation into the time-domain. To formulate it, assume that $X = (X_t)_{t\in\mathbb{Z}}$ is the price process for a single traded asset. Let π_t denote the trading position that an investor takes on this asset at time t. As in the above, we shall use the convention that $\pi_t > 0$ represents a long position (buying the asset), whereas $\pi_t < 0$ represents a short position (selling the asset).

The return of a trading strategy $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_T)'$ over a finite time horizon of T time steps for a realisation $\boldsymbol{x} = (x_1, x_2, \ldots, x_T)'$ of the price process can be written as

$$R_T(\boldsymbol{\pi}|x_0) = \sum_{t=1}^{T} \pi_t (x_0 - x_t) \ . \tag{5.33}$$

In terms of these conventions the expected return μ_S of a trading strategy (conditioned on the initial price x_0) is

$$\mu_S = \langle R(\boldsymbol{\pi}|x_0)\rangle = \sum_{t=1}^{T} \pi_t (x_0 - \mu_t) = x_0 - \boldsymbol{\pi}'\boldsymbol{\mu} \ , \tag{5.34}$$

where we have restricted ourselves in the second step to normalised trading strategies satisfying $\boldsymbol{\pi}'\mathbf{1} = 1$, and where $\mu_t = \langle x_t \rangle$ denotes the expected price at time t.

It is worth remarking at the outset that X could alternatively (and perhaps even more appropriately in the present context) be thought of as the log-price process, in which case $R_T(\boldsymbol{\pi}|x_0)$ would be the log-return of the strategy $\boldsymbol{\pi}$. For the sake of simplicity and definiteness, we shall stick to the language of price processes and returns in what follows.

An optimal trading strategy in the spirit of Markowitz would then be a strategy which minimise the (conditional) variance

$$\mathbb{V}[R_T(\boldsymbol{\pi}|x_0)] = \sum_{t,t'=1}^{T} \pi_t \pi_{t'} \left\langle (x_t - \mu_t)(x_{t'} - \mu_{t'}) \right\rangle = \sum_{t,t'=1}^{T} \pi_t \pi_{t'} \Sigma_{tt'} \ , \tag{5.35}$$

subject to the constraints of normalisation $\boldsymbol{\pi}'\mathbf{1} = 1$ and given mean return $\boldsymbol{\pi}'\boldsymbol{\mu} = x_0 - \mu_S$. In Eq. (5.35), the matrix $\mathbf{\Sigma} = (\Sigma_{tt'})$ now denotes the *auto*-covariance matrix of the price process.

The algebraic side of the problem of finding an optimal trading strategy is now formally fully equivalent to that of finding an optimal portfolio, and the optimal strategy $\boldsymbol{\pi}^*$ takes the form

$$\boldsymbol{\pi}^* = \lambda_1 \mathbf{\Sigma}^{-1}\mathbf{1} + \lambda_2 \mathbf{\Sigma}^{-1}\boldsymbol{\mu} \ , \tag{5.36}$$

with $\mathbf{\Sigma}$ now the *auto*-covariance matrix of the price process rather than the covariance matrix of portfolio returns. Actual values of the Lagrange parameters λ_1 and λ_2 are determined by the constraints as before.

It is well known, and indeed easily verified that the globally optimal solution, which does not impose a restriction concerning the mean return, is compactly given by

$$\boldsymbol{\pi}^*_{\text{GO}} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}} \ . \tag{5.37}$$

As seen before in Eq. (5.36), we have a linear optimisation problem which can be written as

$$\underbrace{\begin{pmatrix} & & & & 1 & \mu_1 \\ & & & & 1 & \mu_2 \\ & & \hat{\mathbf{\Sigma}} & & \vdots & \vdots \\ & & & & 1 & \mu_T \\ 1 & 1 & \dots & 1 & 0 & 0 \\ \mu_1 & \mu_2 & \dots & \mu_T & 0 & 0 \end{pmatrix}}_{\mathbf{A}} \underbrace{\begin{pmatrix} \pi_1^* \\ \pi_2^* \\ \vdots \\ \pi_T^* \\ -\lambda_1/2 \\ -\lambda_2/2 \end{pmatrix}}_{\mathbf{a}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ \mu_p \end{pmatrix}}_{\mathbf{k}} \tag{5.38}$$

where the empirical auto-covariance matrix $\hat{\mathbf{\Sigma}}$ is estimated as Eq. (5.8). The solution is found by $\mathbf{a} = \mathbf{A}^{-1}\mathbf{k}$ and the right hand side can be estimated with ordinary least square regression or an alternative regression method. The global minimum strategy is found with the empirical auto-covariance matrix as

$$\pi^*_{t,\mathrm{GO}} = \frac{\sum_{t'=1}^{N} \hat{\Sigma}^{-1}_{tt'}}{\sum_{t,t'=1} \hat{\Sigma}^{-1}_{tt'}} \,. \tag{5.39}$$

The main problem facing both portfolio optimisation à la Markowitz, and the mean-variance approach to finding optimal trading strategies, is that covariance matrices of portfolio returns or auto-covariance matrices of price processes of traded assets are not known, but need to be *estimated* from empirical market data. The effects of sampling noise in such estimation processes are well studied in the case of portfolio optimisation. As mentioned in the introduction, various strategies to mitigate against such effects, typically guided by random matrix theory, have been investigated in the past.

By contrast, the corresponding random matrix theory for sample auto-covariance matrices that might be invoked for similar purposes for the problem of mean-variance formulations of optimal trading strategies has only recently become available [148]. We shall address the issue of sampling noise in empirical data and the use of spectral theory for the purpose of guiding the choice of "cleaning"-strategies for auto-covariance matrices of market data below in Sec. 5.4. Before that we investigate the effects of sampling noise for some synthetic processes where comparison with known true auto-covariance matrices is possible.

5.3 Results for Synthetic Price Processes

In this section we evaluate the theory developed in the previous section for synthetic price processes. We begin by taking these processes to be either white noise processes or auto-regressive processes of order 1 (see Eq. (5.19)), and then move on to look at the situation where price-*increments* are modelled as white-noise and auto-regressive processes, respectively. For the white noise and auto-regressive price processes, the true auto-covariance matrices are known, and analytical expressions for optimal trading strategies can be given. We then look at the effects of sampling noise, using *estimates* of auto-covariance matrices for various values of the ratio of $\alpha = T/M$ of the length T of the risk horizon (and thus the matrix dimension) and the sample size M used to determine these estimates. The analytical expressions for the true auto-covariance matrices correspond to the $(\alpha \to 0)$-limit in these results.

5.3.1 Synthetic Stationary Price Processes

The primary objectives of the project to analysis a single asset from a portfolio that has been translated to the time domain. In the next few sections we will combine what was learned in the previous two chapters to form the basis of our research. To this end it is assumed that a single assets price is a weakly stationary stochastic signal $\{X_t \mid t \in \{1, 2, ..., N + T\}\}$ where $N, T \in \mathbb{Z}+$. N is the dimension of the empirical auto-covariance matrix and T is the number of realisations. From this signal we calculate its empirical auto-covariance $\hat{\mathbf{\Sigma}}$ and apply optimisation techniques to find the global minimum-variance of the signal $\{X_t\}$, which demonstrates optimal trading strategies π_t^*. We first consider a price process with fluctuations around the trend $\Delta x_t = x_t - \mu_t$ taken to be a Gaussian white noise process, i.e. $\Delta X_t \sim \mathcal{N}(0, \sigma^2)$. The true auto-covariance matrix in this case is proportional to the unit matrix, i.e. $\Sigma_{t,t'} = \sigma^2 \delta_{t,t'}$.

The globally optimal strategy Eq. (5.37) for a time horizon of length T in this case is then readily found to be

$$\pi_{t,\mathrm{GO}}^* = \frac{\sigma^{-2}}{\sum_{t=1}^{T} \sigma^{-2}} = \frac{1}{T} \ . \tag{5.40}$$

Thus, for a white noise process with variance σ^2 the optimal strategy $\boldsymbol{\pi}_{\mathrm{GO}}^* = (1/T, 1/T, ..., 1/T)'$ is uniform over the time horizon T, and independent of the variance of the price process. The analogous result for a Markowitz portfolio of uncorrelated assets is, of course, well known. The optimal liquidation strategy would be to make $N \to \infty$ with $\boldsymbol{\pi} = (1/T, ..., 1/T)'$ and hence the global minimum variance

$$\lim_{T\to\infty} \left\{ (\sigma_p^*)^2 \right\} = \lim_{T\to\infty} \left\{ \tfrac{1}{T} \right\} = 0 \tag{5.41}$$

where σ_p^* is the standard deviation of the optimal trading strategy.

Let us next assume that price fluctuations around the trend are described by an AR(1) process, i.e. an auto-regressive process of order 1 of the form

$$\Delta X_t = a\,\Delta X_{t-1} + \left(\sqrt{1-a^2}\right)\xi_t \ , \tag{5.42}$$

in which $\xi_t \sim \mathcal{N}(0,1)$; for simplicity, we have normalised the process to exhibit fluctuations of variance 1. The parameter a in Eq. (5.42) is required to satisfy $|a| < 1$ for fluctuations to be stationary. The auto-covariance function of this process is known to be given by

$$\gamma(i) = \mathrm{Cov}[\Delta X_t \Delta X_{t-i}] = a^{|i|}. \tag{5.43}$$

The auto-covariance matrix evaluated for a finite time horizon of length T is thus a Toeplitz matrix of the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & a & a^2 & & \cdots & a^{T-1} \\ a & 1 & a & a^2 & & \vdots \\ a^2 & a & 1 & a & \ddots & \\ & a^2 & a & 1 & \ddots & a^2 \\ \vdots & & \ddots & \ddots & \ddots & a \\ a^{T-1} & \cdots & & a^2 & a & 1 \end{pmatrix} . \tag{5.44}$$

Its inverse is a tridiagonal matrix given by

$$\boldsymbol{\Sigma}^{\mathbf{-1}} = \frac{1}{1-a^2} \begin{pmatrix} 1 & -a & 0 & \cdots & \cdots & 0 \\ -a & 1+a^2 & -a & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -a & 1+a^2 & -a \\ 0 & \cdots & \cdots & 0 & -a & 1 \end{pmatrix} . \tag{5.45}$$

The globally optimal strategy Eq. (5.37) for a time horizon of length T in this case is then given by

$$\boldsymbol{\pi}^*_{\mathrm{GO}} = \lambda_1(1, 1-a, \ldots, 1-a, 1)' \ , \tag{5.46}$$

with $\lambda_1 = [2 + (T-2)(1-a)]^{-1}$ fixed by the normalisation-constraint $\boldsymbol{\pi}'\mathbf{1} = 1$. In this case the globally optimal trading strategy turns out to be uniform apart from the two boundary terms. The white noise result is clearly recovered in the $(a \to 0)$-limit for the present result of the AR(1) process, as it should.

Solutions with constraints on the expected return can be given in closed form as well; they are simply obtained by inserting Eq. (5.45) into Eq. (5.36), with Lagrange parameters achieved by solving a pair of linear constraint equations; details will, of course, depend on assumptions concerning the drift, and we refrain from writing them down explicitly.

Fig. 5.1 shows optimal strategies for an AR(1) price process with parameter $a = 0.8$, both for the global optimum as well as for cases with non-zero mean returns imposed. As can be seen from the figure, increasing the expected strategy return from $\mu_S = 4.0 \times 10^{-4}$ to $\mu_S = 1.0 \times 10^{-3}$ changes the optimal strategy Eq. (5.36) from one that is monotone decreasing over the risk-horizon to one which is monotone *increasing*, and starting in fact with a (short-)selling position at the initial time-step $t = 1$.

5.3.2 Synthetic Price Processes with Stationary Increments

The stationarity assumption for the price process used in the previous subsection is clearly unrealistic, and there is obviously need to go beyond that, if the methods discussed in the present investigation are to be useful in practice.

However, once the realm of stationarity is left, some structure is needed on a different level in order to make operational sense of estimating auto-covariance functions and the corresponding auto-covariance matrices defined over a finite time horizon. The structure we shall rely on here is based on the assumption that (fluctuations of) price-process can be described as having *stationary increments*. If one adopts the reading that the processes considered here are actually log-price processes, the assumption of stationarity of their increments is actually a popular assumption in much of Mathematical Finance.

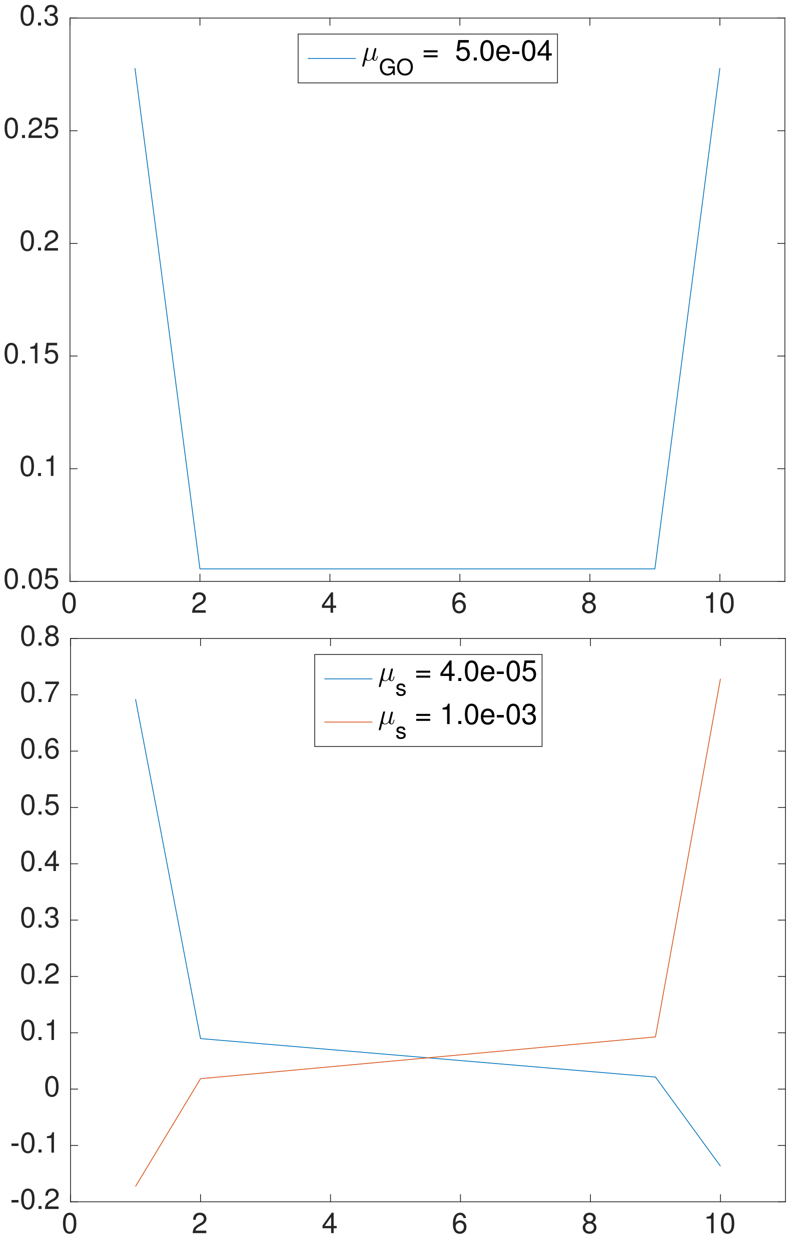


Figure 5.1: Top panel: Globally optimal trading strategy for an AR(1) price process with $a = 0.8$ over a risk horizon of $T = 10$ time steps. Bottom panel: optimal strategies for a process with the same parameter a and a linear drift of the form $\mu_t = 10^{-4}t$, imposing expected strategy returns of $\mu_S = 4 \times 10^{-5}$ (blue solid line) and $\mu_S = 1 \times 10^{-3}$ (solid orange line).

In what follows we assume that the (log-) price process $X = (X_t)$ exhibits stationary increments, i.e. that

$$X_t = X_{t-1} + Y_t \tag{5.47}$$

with $Y_t = \langle Y_t \rangle + \Delta Y_t = \mu_t - \mu_{t-1} + \Delta Y_t$ with zero-mean fluctuations ΔY_t. In terms of these conventions, we can write the return of a strategy $\boldsymbol{\pi} = (\pi_t)$ for a given realisation $\boldsymbol{x}$ as

$$R_T(\boldsymbol{\pi}) = \sum_{t=1}^{T} \pi_t (x_0 - x_t) = \sum_{t=1}^{T} \pi_t \Big[(\mu_0 - \mu_t) - \sum_{\tau=1}^{t} \Delta y_\tau \Big]. \tag{5.48}$$

The expected return is given by the first contribution on the r.h.s, while the variance is

$$\mathbb{V}[R_T(\boldsymbol{\pi})] = \sum_{t,t'=1}^{T} \pi_t \pi_{t'} \Big[\sum_{\tau=1}^{t} \sum_{\tau'=1}^{t'} \langle \Delta y_\tau \Delta y_{\tau'} \rangle \Big]. \tag{5.49}$$

This is of the same structure as Eq. (5.35), with the auto-covariance matrix $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}^X = (\Sigma^X_{t,t'})$ of the non-stationary price process expressed in terms of the auto-covariance matrix $\boldsymbol{\Sigma}^Y = (\Sigma^Y_{t,t'})$ of the process of price increments as

$$\Sigma^X_{t,t'} = \sum_{\tau=1}^{t} \sum_{\tau'=1}^{t'} \langle \Delta y_\tau \Delta y_{\tau'} \rangle = \sum_{\tau=1}^{t} \sum_{\tau'=1}^{t'} \Sigma^Y_{\tau,\tau'} \ . \tag{5.50}$$

This relation between the auto-covariance matrices of process and the corresponding process of increments can be compactly expressed in matrix form as

$$\boldsymbol{\Sigma}^X = \boldsymbol{P} \boldsymbol{\Sigma}^Y \boldsymbol{P}' \ , \tag{5.51}$$

where $\boldsymbol{P}$ is a lower triangular constant matrix of ones,

$$\boldsymbol{P} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} . \tag{5.52}$$

The mean-variance approach to strategy optimisation then yields optimal trading strategies of the form Eq. (5.36), with the auto-covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}^X$ of the price process expressed in terms of the auto-covariance matrix $\mathbf{\Sigma}^Y$ of the process of stationary increments according to Eq. (5.51).

Taking the price increments to be a white noise process $\Delta Y_t \sim \mathcal{N}(0, \sigma^2)$, we have $\Sigma^Y_{t,t'} = \sigma^2 \delta_{t,t'}$ so $\mathbf{\Sigma}^{-1} = \sigma^{-2}(\boldsymbol{PP'})^{-1}$, where $(\boldsymbol{PP'})^{-1}$ is found to be of tridiagonal form,

$$(\boldsymbol{PP'})^{-1} = \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & \dots & & -1 & 1 \end{pmatrix} . \tag{5.53}$$

The globally optimal strategy Eq. (5.37) in this case is then simply

$$\boldsymbol{\pi}^*_{\mathrm{GO}} = (1, 0, 0, \dots 0)' \ , \tag{5.54}$$

i.e., it consists of taking a single long position at the initial time step.

If we assume an AR(1) process, of the form Eq. (5.42), for the fluctuations of the price increments, i.e.

$$\Delta Y_t = a\,\Delta Y_{t-1} + \left(\sqrt{1-a^2}\right)\xi_t \ , \tag{5.55}$$

then it is $\mathbf{\Sigma}^Y$ which is given by Eq. (5.44); it turns out that $\mathbf{\Sigma}^{-1} = (\boldsymbol{P}\mathbf{\Sigma}^Y\boldsymbol{P'})^{-1}$, too, can be evaluated in closed form, giving

$$\boldsymbol{\Sigma}^{\mathbf{-1}} = \frac{1}{1-a^2}\begin{pmatrix} C & -A^2 & a & 0 & \cdots & \cdots & & \cdots & 0 \\ -A^2 & 2B & -A^2 & a & 0 & & & & \vdots \\ a & -A^2 & 2B & -A^2 & a & 0 & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & & & & & & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 0 & a & -A^2 & 2B & -A^2 & a \\ \vdots & & & & 0 & a & -A^2 & C & -A \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & a & -A & 1 \end{pmatrix} .$$

in which we use the abbreviations $A = 1 + a$, $B = 1 + aA$ and $C = 1 + A^2$.

In this case the globally optimal strategy Eq. (5.37) is of the form

$$\boldsymbol{\pi}^*_{\mathrm{GO}} = (1 + a, -a, 0, \ldots, 0)' \ , \tag{5.56}$$

i.e. it consists of taking a single long position at the first time-step, which is then partially offset by a short position at the second time step if $a > 0$, whereas it is followed by a further long position if successive price increments are *anti*-correlated ($a < 0$). Note that the solution for white noise increments is correctly recovered as the ($a \to 0$)-limit of the AR(1) results.

Once more, solutions with constraints on expected returns can be given in closed form; in analogy to the procedure described for the case of stationary price processes, they are obtained by inserting Eq. (5.45) into Eq. (5.36), with Lagrange parameters obtained by solving a pair of linear constraint-equations.

We find, and shall demonstrate below, that the procedure predicts non-trivial changes of strategy as constraints on expected returns are varied. Once more, details will depend on assumptions concerning the drift, and we refrain from producing explicit equations here. We will report our analytical results alongside numerical results which take sampling errors arising from finite sample fluctuations on *estimated* auto-covariance matrices into account

5.3.3 Sampling and The Effects of Noise

Having analytical results for synthetic price processes available allows one to estimate the effects of sampling noise on optimal strategies and on risk-return profiles. In practice, the analytic structure of an underlying price process will not be known, and auto-covariance matrices will have to be estimated on the basis of finite samples, i.e. the design of optimal strategies will have to be based on *sample auto-covariance matrices* $\hat{\mathbf{\Sigma}}$.

For a stationary price process, samples taken along a realisation of the process can be taken to define the elements of $\hat{\mathbf{\Sigma}}$ via

$$\hat{\Sigma}_{t,t'} = \frac{1}{M-1} \sum_{k=1}^{M} \Delta x_{t+k} \Delta x_{t'+k} \ . \tag{5.57}$$

This procedure introduces sampling noise; estimated auto-covariance matrix elements $\hat{\Sigma}_{t,t'}$ will exhibit $\mathcal{O}(M^{-1/2})$ fluctuations about their corresponding true counterparts $\Sigma_{t,t'}$, which is observed from the central limit theorem. When assessing the effects of sampling noise via the influence on spectra, one expects the relevant parameter to be the aspect ratio $\alpha = T/M$, i.e. the ratio of the number of time-lags considered and the sample-size used to estimate matrix elements. We shall use this parameter in what follows to parametrise the influence of sampling noise, with the ($\alpha \to 0$)-limit corresponding to the situation without sampling noise, i.e. with true asymptotic auto-covariances known.

If the price process is not stationary, but has stationary increments, one can use Eq. (5.50) and Eq. (5.51) to express the auto-covariance matrix $\mathbf{\Sigma}^X$ of the price process in terms of the auto-covariance matrix $\mathbf{\Sigma}^Y$ of the process of price increments. For the latter it is legitimate to use an estimator by sampling along a realisation, so one can define $\hat{\mathbf{\Sigma}}^Y$ via

$$\hat{\Sigma}^Y_{t,t'} = \frac{1}{M-1} \sum_{k=1}^{M} \Delta y_{t+k} \Delta y_{t'+k} \tag{5.58}$$

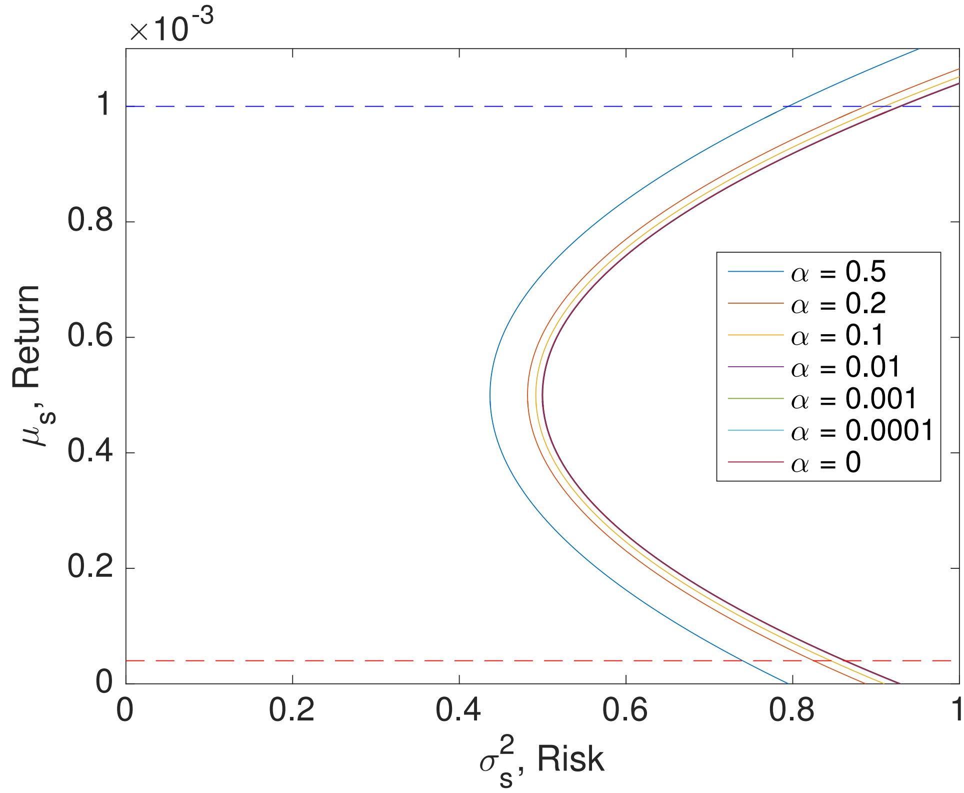


Figure 5.2: Risk-return profile for an AR(1) price process with the same parameters as in Fig. 5.4 for various levels of sampling noise parameterised by α. Results are obtained by averaging over 10^7 samples as in Fig. 5.3. Note in particular that sampling noise leads to an *under-estimation of risk*. The two horizontal dashed lines indicate two values of the target return for which optimal trading strategies are reported in Fig. 5.4 below.

and

$$\hat{\mathbf{\Sigma}}^X = \boldsymbol{P}\hat{\mathbf{\Sigma}}^Y\boldsymbol{P}' \ . \tag{5.59}$$

In Fig. 5.2 we show the risk-return profile for the case of an AR(1) price process for various aspect ratios α, ranging from $\alpha = 0.5$ down to $\alpha = 10^{-4}$, with the noise-free case $\alpha = 0$ also included. Note that sampling noise leads to a systematic underestimation of risk, though results quickly approach the noise-free limit as α becomes small.

Fig. 5.3 exhibits the weights of the globally optimal (minimum risk) trading strategy for this process, while Fig. 5.4 gives weights of optimal trading strategies for two different values of the target return (indicated by the two horizontal dashed lines in Fig. 5.2. In this case we assume a small drift $\mu_t = 10^{-4}t$ of the underlying price process. It is noticeable that an increase in the required target return leads to a qualitative change of the optimal strategy, with the larger target return requiring to take an initial short position at the beginning of the trading period.

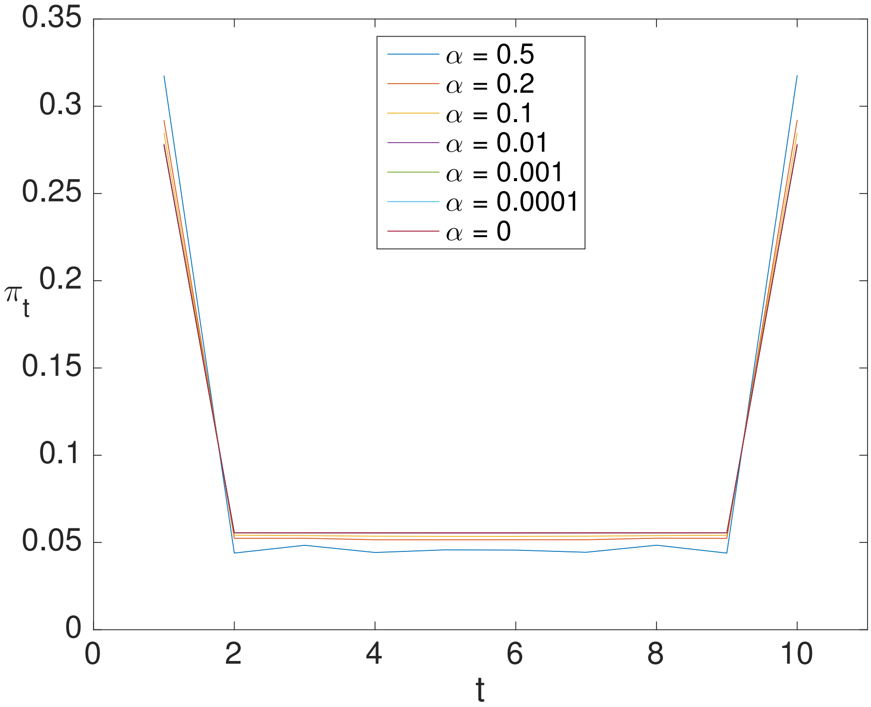


Figure 5.3: Globally optimal trading strategies for an AR(1) price process with $a = 0.8$ over a risk horizon of $T = 10$ time steps, using *estimated* auto-covariance matrices. Data are shown for various values of the ratio $\alpha = T/M$ of risk horizon and sample size M used to estimate auto-covariances according to Eq. (5.57): optimal strategies (with solid lines as guides to the eye) are obtained by averaging over 10^7 samples. Standard deviations are also shown; they rapidly decrease with α – not shown here as it obscures the shape of the strategy. Results obtained for the *true* auto-covariance function, the ($\alpha \to 0$)-limit, are included for comparison. Note that average strategies obtained for finite samples are very close to the $\alpha = 0$ results.

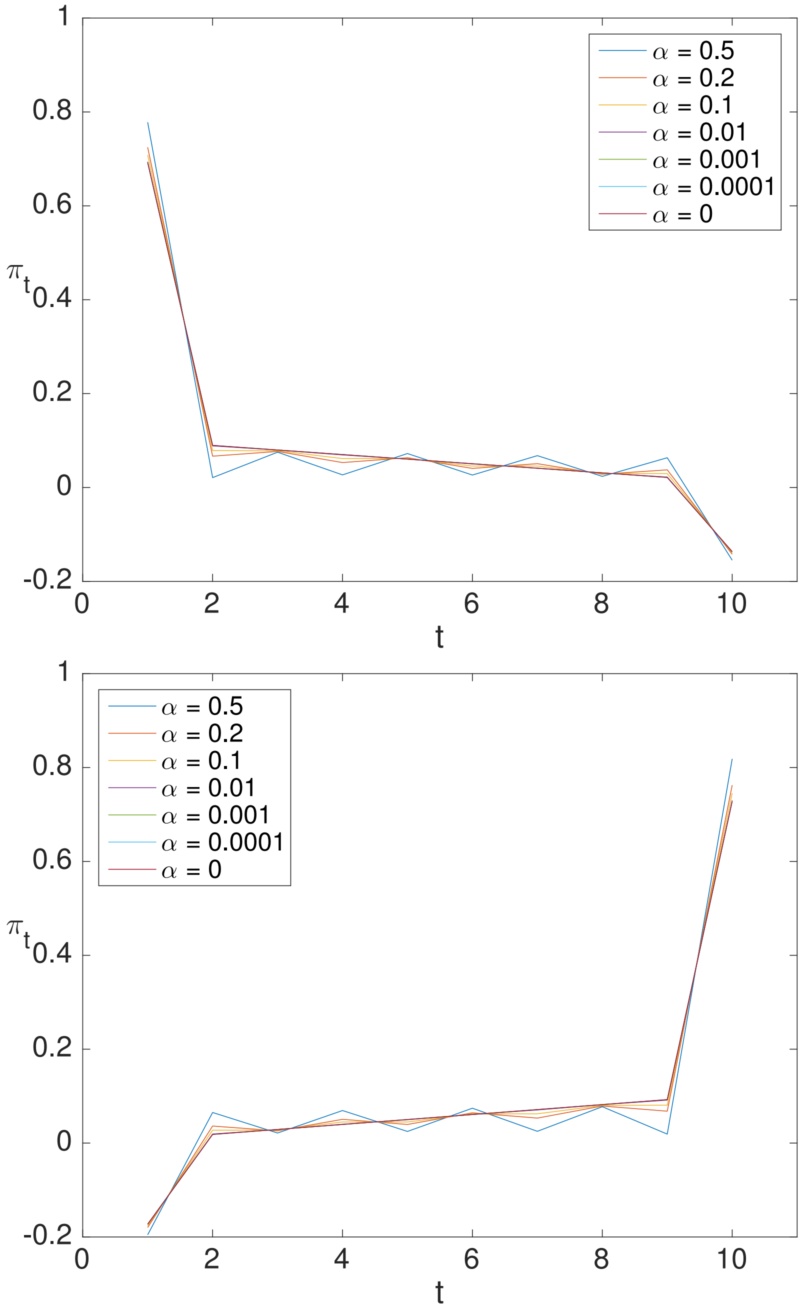


Figure 5.4: Top panel: Optimal strategies for an AR(1) process with $a = 0.8$ and a linear drift of the form $\mu_t = 10^{-4}t$ as in Fig. 5.1, with imposed expected strategy return of $\mu_S = 4 \times 10^{-5}$. Shown are average trading strategies for various levels of sampling noise parameterised by non-zero α, obtained by averaging over 10^7 samples. Average results are close to those obtained using true asymptotic autocovariance matrices in the ($\alpha \to 0$)-limit, which are included for comparison. Bottom panel: optimal trading strategies for an AR(1) process with the same parameters as in the left panel, but now with $\mu_S = 1 \times 10^{-3}$.

Turning to the situation where we use an auto-regressive process to describe the statistics of price *increments*, we see from a comparison of Fig. 5.5 and Fig. 5.2 that risk levels are significantly larger compared to the situation where the same underlying process describes the fluctuations of the price process itself.

This concludes our collection of results for synthetic price processes, where the underlying true auto-covariances are known. We now turn to applying the framework to empirical data, where this is not the case.

5.4 Empirical Data

In what follows we apply our framework to empirical data, using daily adjusted close data of the S&P500, spanning the period 03 Jan 1950 to 20 Apr 2015.

This is perhaps the point to notice that we are not advocating that using the variance of trading strategy returns constitutes the best way of capturing risk in real market data. Indeed, given that market returns are known to have fat-tailed distributions, variance can at best be regarded as a proxy for risk. However, our primary goal here is not to explore a wider family of possible risk measures, but rather to define a reformulation of the popular mean-variance optimisation strategy in the time domain, and to begin investigating its properties.

5.4.1 The Spectrum of the S&P500 Auto-Correlation Matrix

Before turning to the evaluation of optimal trading strategies and risk-return profiles we shall have a look at the spectrum of the auto-covariance matrices of the data, taking time windows of $T = 50$, and sample sizes of $M = 100$, hence $\alpha = 0.5$. Auto-covariance matrices of the price process are obtained as described in Sec. 5.3.3, by first evaluating auto-covariances of the return process, assuming stationarity across individual sample-windows. In order to obtain meaningful statistics across the entire data set, we transform the return series in each time window to exhibit unit-variance increments, and then obtain auto-covariances of the thus normalised price process using the transformation Eq. (5.59).

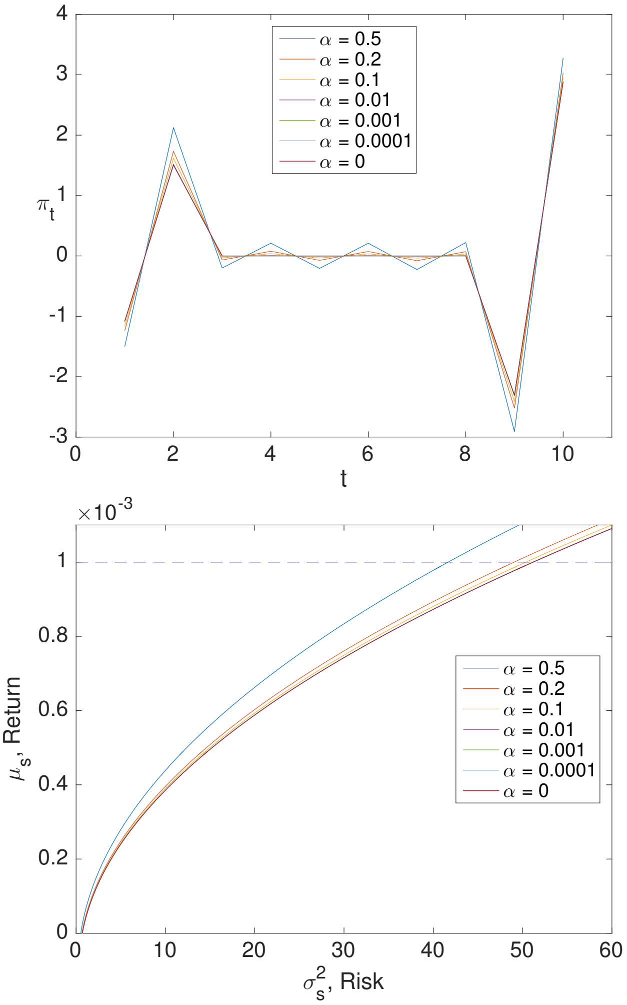


Figure 5.5: Top Panel: Optimal strategies for a setup where the fluctuations of the *price-increments* are described by an AR(1) process with $a = 0.8$; a linear drift of the form $\mu_t = 10^{-4}t$ is assumed for the price process, and an expected strategy return of $\mu_S = 1 \times 10^{-3}$ is imposed. Shown are average trading strategies (solid lines) obtained by averaging over 10^7 samples for various levels of sampling noise parameterised by non-zero α. Average results are close to those obtained using true asymptotic auto-covariance matrices in the ($\alpha \to 0$)-limit, which are included for comparison. Bottom Panel: risk-return profile for this setup, with the horizontal dashed line indicating the expected strategy return imposed in the data of the top panel. The bottom panel should be compared with Fig. 5.2, which exhibits the risk return profile for an AR-1 price process.

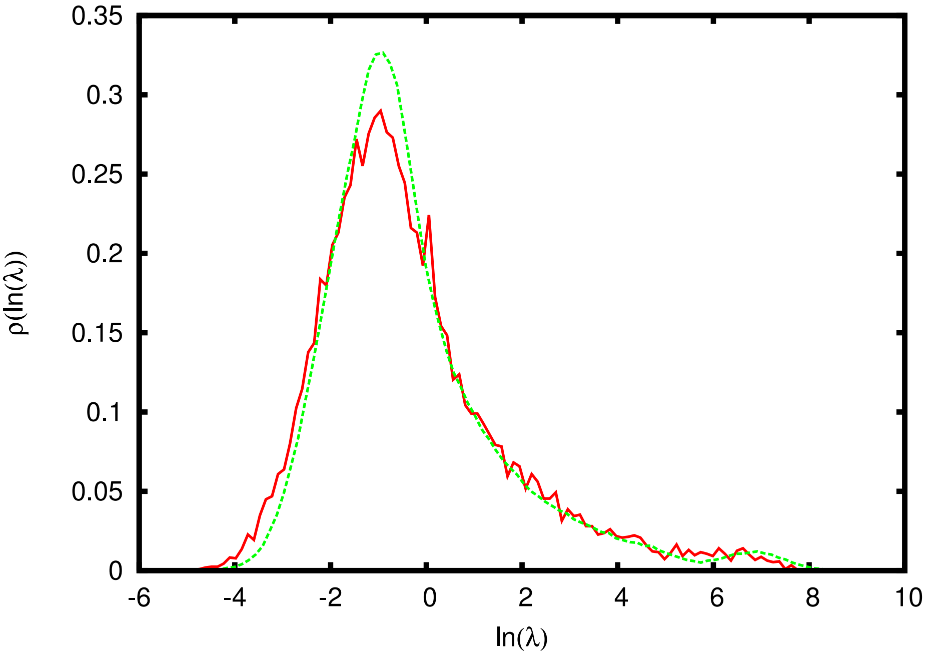


Figure 5.6: Spectrum of the sample auto-covariance matrix of the S&P500, normalised as described in the main text, using $T = 50$ time lags and an aspect ratio $\alpha = 0.5$, i.e. samples of size $M = 100$ to define the auto-covariances (red full line). Also shown is a comparison with the spectrum of an auto-covariance matrix for a price process with *independent* unit variance increments (green dashed line). The two are remarkably close.

As can be seen in Fig. 5.6, where we plot the density of logarithms of eigenvalues for $\mathbf{\Sigma}^Y$, the spectrum is *very broad*, spanning several orders of magnitude. For comparison we include the spectrum for a process with *independent* unit variance increments using the same values of T and M, and we notice that the two are remarkably close. This is not completely unanticipated, as it is one of the widely reported 'stylised facts' in the field that return-series have very short correlation-times. We will use this type of spectral comparison below to inform the auto-covariance matrix cleaning strategy that we will use for the purpose of noise reduction.

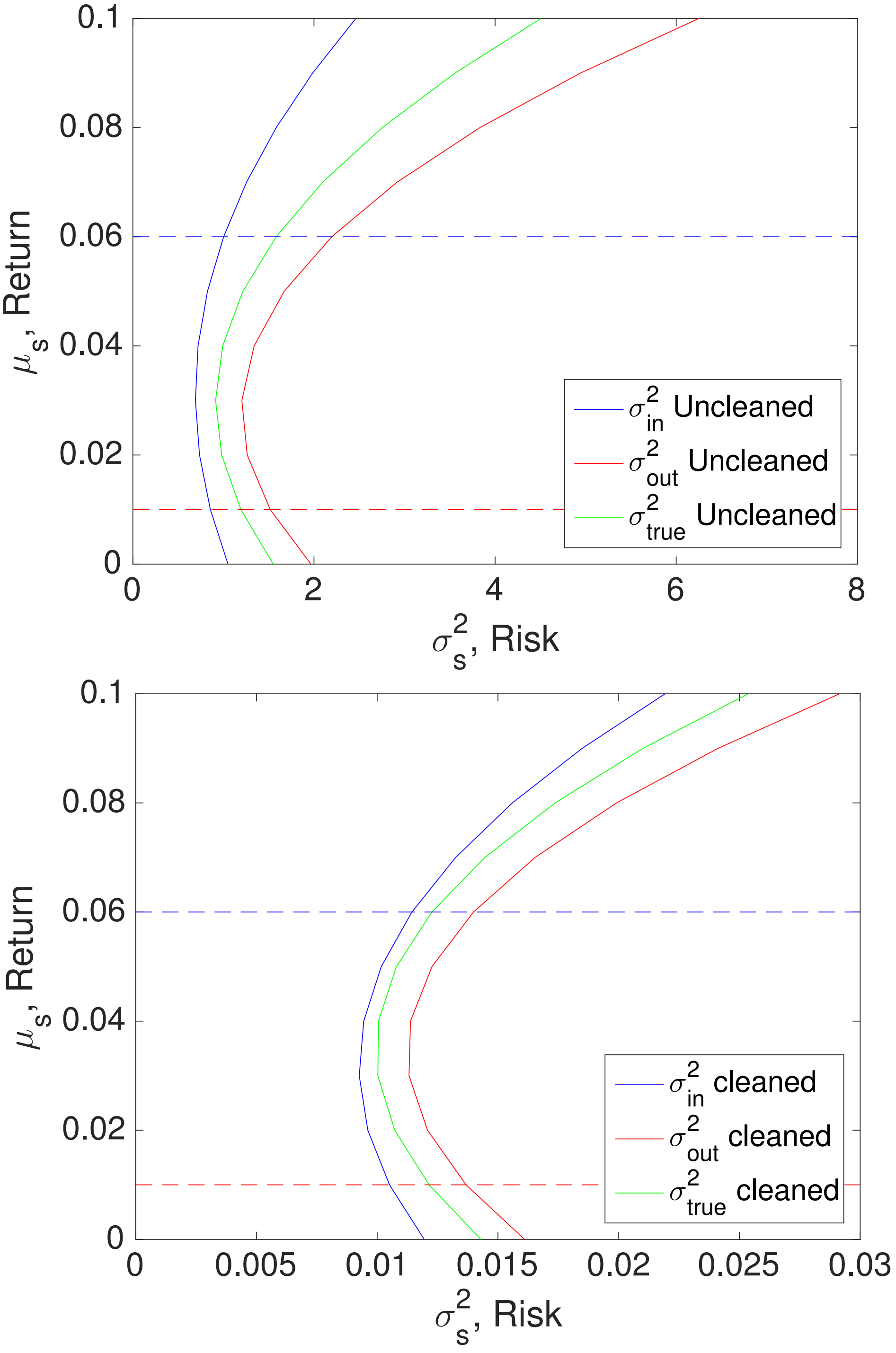


Figure 5.7: Risk-return profile of optimal trading strategies on the S&P500 data. Left: risk-return profile obtained from measured auto-covariance matrices. Right: risk-return profile obtained using cleaned versions of auto-covariance matrices. Horizontal dashed lines denote target strategy returns μ_S for which optimal strategies are reported in Fig. 5.8.

5.4.2 Optimal Trading Strategies and Auto-Covariance Matrix Cleaning

In Fig 5.7 we report the risk-return characteristics for optimal trading strategies on the S&P500, using sample-auto-covariance matrices of $T = 50$ time lags, and sample size $M = 100$ as in Fig. 5.6. We report results obtained for auto-covariance matrices, as measured via Eq. (5.58) and Eq. (5.59), and compare them with results obtained by applying a cleaning strategy to these, which we shall describe below. We use *realised* returns defined by linear trends in each data window to compute risk-return profiles, and use conventions for in-sample risk, true risk and and out-of-sample risk as in [159], taking the average auto-correlation matrix across the entire time series as a proxy for the true auto-correlation. The convention from [159] for risk are defined as:

1. In-sample risk:
$$\sigma^2_{\text{in}} = \boldsymbol{\pi}'_{\text{E}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\pi}_{\text{E}} \tag{5.60}$$

2. True risk:
$$\sigma^2_{\text{true}} = \boldsymbol{\pi}'_{\text{C}}\boldsymbol{\Sigma}\boldsymbol{\pi}_{\text{C}} \tag{5.61}$$

3. Out-of-sample risk:
$$\sigma^2_{\text{out}} = \boldsymbol{\pi}'_{\text{E}}\boldsymbol{\Sigma}\boldsymbol{\pi}_{\text{E}} \tag{5.62}$$

The subscript E denotes an empirical estimations and subscript C is an analytic calculation. Note, that the reduction of risk that can be obtained through cleaning is substantial. Fig. 5.8 exhibits optimal trading strategies for the S&P500, showing both the minimal risk solution and risk-optimal solutions for two different non-zero target strategy returns. Apart from the effect of reducing risk, we find that the effect of cleaning is also to create strategies that are “smoother” than those obtained without cleaning.

Let us finally turn to the cleaning approach that is used to get the data described above. In the context of *covariance* matrices of financial data, strong similarities were observed between empirical correlation matrix spectra and the Marčenko-Pastur law expected for high-dimensional uncorrelated data. One of the cleaning strategies that has been suggested due to such similarities is referred to as 'clipping' [133, 159]. It analyses correlation matrices by performing a spectral decomposition, and regards the bulk of a sample correlation matrix spectrum, which resembles the Marčenko-Pastur law, as noise. It then transforms correlation matrices by keeping large eigenvalues outside the majority, and replacing those in the bulk by their average, thereby avoiding small eigenvalues in the modified matrix. In the present case, the phenomenology is rather different; there are no eigenvalues of the (normalised) sample auto-covariance matrices that can be regarded as lying significantly outside the bulk of the spectrum predicted for uncorrelated increments. So there would be no clear guidance coming from random matrix theory that could form the basis of a clipping-type procedure.

We, therefore, decided to apply a 'shrinkage' method to our data. To the best of our knowledge, this procedure was first proposed by Stein [160], and has recently found renewed interest in the Mathematical Statistics [136, 142] and Econophysics [161] communities.

Based on the observation reported in Fig 5.6 that the (normalised) auto-covariance spectra of the S&P500 and of a synthetic process with independent increments are indeed rather similar, we apply the shrinkage procedure to the sample auto-covariance matrixes of the S&P500 increments $\hat{\mathbf{\Sigma}}^Y$, shrinking them towards a target matrix $\mathbf{D}$ given by the diagonal matrix of *variances* of the increments (which would indeed describe a process of independent increments), i.e. towards $\mathbf{D} = \mathrm{diag}(\{\hat{\Sigma}_{t,t}\})$, using the substitution rule

$$\hat{\mathbf{\Sigma}}^Y \leftarrow \rho\mathbf{D} + (1-\rho)\hat{\mathbf{\Sigma}}^Y \ , \tag{5.63}$$

and transforming the shrunk $\hat{\mathbf{\Sigma}}^Y$ thus obtained to define the cleaned estimate of $\hat{\mathbf{\Sigma}}^X$ using the transformation Eq. (5.51). The proper value for the parameter ρ in this procedure is determined from the data as described in [136, 142].

5.5 Summary and Discussion

To summarise, in the present study we have a reformulation of Markowitz' mean-variance optimisation in the time domain to obtain optimal trading strategies for a single traded asset over a finite discrete time horizon. Using simple linear algebra, one gets such optimal trading strategies as sequences of buy, hold, and sell instructions for that asset, which minimise the market fluctuations of the return generated by this sequence of instructions over a given time horizon, subject to suitable constraints. The procedure requires the auto-covariance matrix of the price process (and estimates for expected prices) during the risk horizon as input.

We investigated this problem for some synthetic price processes, taken to be either second order stationary or be described by second order stationary increments. Analytic expressions are given for the cases where the price and the return processes are described by *i.i.d.* or by auto-regressive fluctuations.

We compare analytic solutions with numerical results for situations where auto-covariance matrices have to be estimated from finite samples, which is the situation typically encountered in practice. For the synthetic processes for which true auto-covariance matrices are known the effects of sampling noise on optimal strategies and on risk-return profiles can thus be quantitatively assessed. We find that in general sampling noise leads to an underestimation of risk, but that asymptotic results are well approximated when samples used to estimate auto-covariance matrices are sufficiently large. A ratio $\alpha = T/M < 0.1$, i.e. sample sizes ten times the length of the risk-horizon appears to be desirable from this point-of-view.

From the financial point of view on the other hand, it is always desirable to use time series as short as possible for estimation, to avoid letting (possibly) outdated data influence current trading strategies. Small samples, however, increase the effects of sampling noise, and it is for this reason that cleaning strategies have an important role to play. Looking at the S&P500 data, we found that (normalised) auto-covariance spectra closely resemble those one would expect for price processes with independent increments, and it is this observation that motivates our choice of target matrix within a shrinkage cleaning strategy.

We observe that the auto-covariance matrix cleaning gives rise to smoother trading strategies and that it also leads to a reduction of risk in risk-return profiles.

A natural generalisation of the present work would deal with a multi-period multi-asset version of a mean-variance formulation of optimal trading strategies. One approach to this problem is to redefine the covariance matrix in the optimisation problem to a block matrix. This block matrix will have its diagonal matrices as the auto-covariance matrices for the N assets and the off-diagonal as the cross covariance matrices. While some work has been done in this direction in the past (see. e.g. [147] and references therein) the solution presented in [147] remains somewhat formal and restricted to the case without correlations in time.

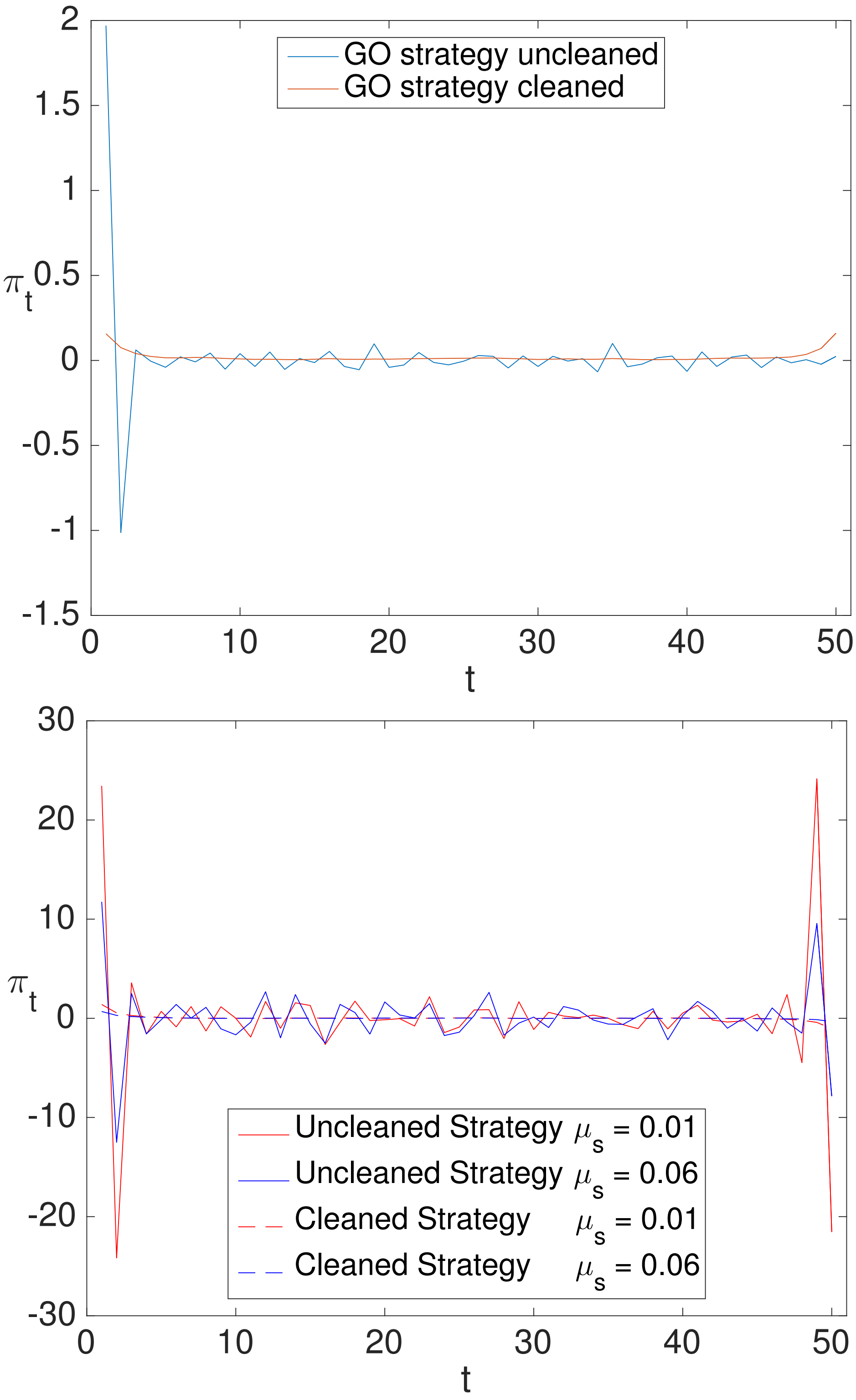


Figure 5.8: Top Panel: globally optimal strategy for the S&P500, showing both results before and after cleaning. Bottom Panel: optimal strategies for the two target returns of $\mu_S = 0.01$ and 0.06 indicated in Fig. 5.7 .

Chapter 6

Conclusion

This chapter reviews the models developed in Chapters 2-5, with the addition of possible future research directions. The goal of the thesis was to explore different ways of modelling prices in two different types of markets: financial and gambling. This was achieved from a number of different angles using statistics, and stochastic and time series models.

In Chapter 2, we discussed the service sector known as the *online gambling industry*. This market has grown substantially in the last decade and has been driven, partly, by the growth of the internet.

Sec. 2.1 discusses the main features of the gambling market and shows that the online sector is the fastest growing part. It was crucial in this section to illustrate the contrast between gambling and investment – because one can be perplexed by their similarities. These similarities can be quite striking, to the point if one explores the job market one will now find gambling jobs advertised with titles such as *Quant* or *Algo-trader*; roles usually reserved solely for the finance world. This resemblance between the financial profession and gambling profession has materialised through the similarity in the skill sets needed to model such entities as the *order book* and the different guises of data science. One must be very clear; the differences emerge from the fundamental role that investment plays in society, as it provides society with the means to prosperity and progress; such as faster computers and funding for research degrees. Whereas, the gambling industry is a constituent of the services sector – more specifically the entertainment sector.

Using the data provided to the author by [3], we were able to perform an extensive study of the in-play horse racing markets. We gave details of this particular market in Chapter 2 and explained how it is constructed through components such as race selection, horse selection, order-book, backing and laying odds. We showed that exchanged assets in gambling markets are odds and illustrated how they are auctioned, similar to financial market order-books but with different quantities such as volume being the quantised in terms of a numéraire – in Betfair's case Great British Pounds Sterling. The decimal odds have a dual meaning in the gambling markets, as this is the instrument that is exchanged but it also has the interpretation of value; which makes it similar to price. We found that bets can be hedged with a two position process where one backs and then lays, or vice versa, and gains (or losses) are extracted from the change in the odds between the two positions.

Sec. 2.2 we explain the various fields that constitute the in-play horse racing data set with a detailed description and discussion. The statistical properties are explored in detail for the in-play market signals in Sec. 2.3. We present sample statistics of the data set such as the average market liquidity in Table 2.2. The Table 2.3 depicts the estimation of the sample moments for the increments of the last price matched signals which indicate that the distributions are not normal and are evidently leptokurtic. We discovered some interesting statistics, shown in the Table 2.4, where we estimated the sample average for the initial, last price matched for the favourite and long-shot; and observed that the mean value corresponds to the unconditional probability of winning the race, estimated as Eq. (2.16). By applying statistical dispersion and entropic measures to the last price matched signals we identify that on average the dispersion in the odds increases in time, see Fig. 2.6.

In Sec. 2.4 we find a significant similarity between the order statistics of the randomly broken stick, of unit length, and the statistical market estimation of initial odds. We observe that the empirical values of the implied odds and true winning probabilities come close in value. We conclude – partly because of the results in Table 2.5 – that this betting market is **initially** to some degree informationally efficient. By informationally efficient we mean the market on average can correctly order the selections, but for each race, this is comparable to randomly splitting the unit interval as explained previously. Discrepancies in the market's estimation of the long-shot odds (or similar low-rank selections), implies that the market is inefficient at accurately ranking the selections, especially when the total number of selections is large. We further expand this prediction by making the assumption that the implied odds reflect the horses' (and jockeys') "ability" to win the race. This term "ability" is in some way vague but we stipulate the semantics here that this refers to the true probability of the horse (and jockey) winning the race, such as the exponential distribution defined in Eq. (2.22).

Chapter 3, describes two models: a novel toy model in Sec. 3.2 and a traditional statistical arbitrage trading strategy known as pairs in Sec. 3.3.

The novel toy model makes the assumption that the competing horses **in-play** have a relative position (or performance) that is a local martingale and is defined through a set of Itô processes. Making this assumption means we are treating the market in accord with the *Efficient Market Hypothesis* implying that the market is in a state of perfect competition, and thus no one market agent can manipulate the prices to their advantage. Using these relative positions one can calculate the probability of the horse winning the race by a numerical procedure where another distribution is centred on the current position of each horse and sampled from. One can see from Fig. 3.1 that we are estimating (numerically) the probabilities of a particular horse being in the lead in the next time step, the blue area in Fig. 3.1. Fig. 3.2 shows that the odds exhibit a cointegrating behaviour between competing horses and the probabilities converge to the correct values: winner → 1 and loser → 0. An issue concerning this model is the constant over-round, which is always one. It would be interesting, as a line of further investigation, to examine the possibility of creating a similar model but with a varying over-round.

The traditional statistical arbitrage trading strategy known as pairs is explained from a financial point-of-view in Sec. 3.3. This trading strategy is then reverse engineered in Sec. 3.4 and applied to in-play horse racing data. The results of the pairs trading algorithm are shown in Figs. 3.4 and 3.5. We see from Fig. 3.4 that the algorithm uniformly sets up betting positions across the time window. The average rate-of-return is displayed along with the average Sharpe-ratio in Fig. 3.5. The results are shown in Fig. 3.5 should be treated with a high degree of scepticism as a final rate-of-return $\approx 370\%$ is suspiciously large and unlikely to be attained in reality. Two reasons come to mind why rate-of-return is so high: (1) there is no price impact when trading and (2) that backs and lays are always matched and at the front of the queue. It would be interesting to explore these two deficiencies in the model by calibrating a price impact model as found in [162] and developing order book dynamics to account for queues such as [125].

In Chapter 4, we gave a detailed review of a body of work known as *Information Based Finance* or the *Brody-Hughston-Macrina* (BHM) framework, found in the following literature [86–110]. The bulk of the review is given in Secs. 4.1 and 4.2 deriving the pricing model, see Eq. (4.25), which is the basis of our model for trading in Sec. 4.3.

Sec. 4.3 the BHM framework is used to develop a synthetic agent based market where under certain conditions the agents trade. To achieve this end, we created a mechanism for trading to take place in Sec. 4.3, with the addition of a specification of how the market agents' price/information is updated after trading. The different market behaviour was created through the categorisation of agents into the following: market makers, Sec. 4.3.3; informed traders, Sec. 4.3.4; and noise traders, Sec. 4.3.5. The agent was categorised in this way to give the market and each type of agent a different trading mechanism, which in turn led to various market price features. To help highlight the different features generated by the different market agents we introduced the inventory control mechanism. The inventory control was defined in Eq. (4.91), and is a utility function that controls the amount an agent buys and sells relative to their absolute value in inventory. When applying Eq. (4.91) one can observe the features generated by the informed traders and noise traders on the market price from.

Sec. 4.4 presents the results for the trading model where we display the distributions for the market returns and first passage times for trading for the following market configurations $(2, 0, 0)$, $(2, 1, 0)$ and $(2, 0, 1)$, this notation represents (N_{MM}, N_{IT}, N_{NT}) where N_{MM} is the number of marker makers, N_{IT} is the number of informed traders, and N_{NT} is the number of noise traders.

For the configuration $(2, 0, 0)$ the half spread parameter δ, which is half the distance between the market making agents buy and sell price, is varied. We found that when the half spread δ is increasing the overall frequency of trades reduces and the standard deviation of the returns is growing. The reason for these two effects is that increasing the half spread increases the amount agents have to diffuse away from each other to trade, which in turn increases the standard deviation of the returns and the time waited in between trades by agents; this also reduces the overall number of trades by the agents. We observed that when the inventory control was applied, the number of overall trades reduced to a constant number, as did the standard-deviation, and the kurtosis was found to have increased.

For the configuration $(2, 1, 0)$ the half spread of the informed trader δ_I, which is half the distance between the informed trader agent's buy and sell price, is varied while the market makers' half spread is held constant at δ. We found that when the half spread δ_I is increasing the overall frequency of trades reduces and the standard deviation of the returns is growing. When the inventory control was applied we observed – the tentative result of – a bimodal distribution (see the bottom plot in Fig. 4.11).

For the configuration $(2, 0, 1)$ the threshold parameter of the noise trader γ, which is the value sampled from a uniform distribution has to exceed to trade, is varied while the market markers half spread is held constant at δ. We found that this had no effect on the overall number of trades and standard-deviation both of which remain constant. The reason for this is that trading is dominated by the market makers who have a small half spread which does not change. When inventory control is applied we observed that the number of trades and standard-deviation is decreasing as γ increases which is because the larger this parameter is, the smaller the probability becomes for a noise trade to occur and thus the longer one has to wait for a noise trade.

Sec. 4.6 introduces a non-linear rate function in the BHM information process, instead of a linear function. The new non-linear Eq. (4.107) was defined in terms of a general function $g(t)$ that obeys the condition Eq. (4.70). Using this non-linear version of the information process we apply Itô's Lemma and derive a new dynamical price process Fig. 4.131) in terms of $g(t)$. This non-linear version of the information process has a similar structure to the BHM price model but is more general.

In Sec. 4.7 the original BHM theory was fitted to the winning horse signals from the data set. It was found that the average last price matched could be modelled with the original BHM model using an information rate parameter $\hat{\sigma}$ that depends on course distance according to a power law with a positive exponent, see Fig. 4.20. This result gives the interpretation that the odds in short races are governed more by noise than is found in longer races. Sec. 4.7.2 took the non-linear BHM model, as derived in Sec. 4.6, and fitted it to the same data set. This estimation gave a different result to when the original BHM model was fitted as the information rate function $\hat{\sigma}(t)$ that was fitted to the non-linear BHM can be regarded as either piecewise linear or a polynomial of order 6. There was an issue found with this result as one can see that in Fig. 4.22 when $t \to T$ the function does not satisfy $\hat{g}(t) \to 1$, which violates the condition Eq. (4.70). Therefore, to further improve this fit we would have to investigate better ways of fitting this model; perhaps with regularisation such as the ridge or lasso regression [163], both of which penalise the dimension of the regression to prevent overfitting a model.

Chapter 5, in its essence gave a reformulation of Markowitz' mean-variance optimisation in the time domain to obtain optimal trading strategies for a single traded asset over a finite discrete time horizon, discussed in Sec. 5.2.1. We obtained such optimal trading strategies as sequences of buy, hold, or sell instructions for that asset, which minimise the market fluctuations of the return generated by this sequence of instructions over a given time horizon, subject to suitable constraints. The methodology required the empirical estimation of the auto-covariance matrix of a given price process, and estimates for expected prices, during the risk horizon as input.

We applied this framework to a set of synthetic price processes. Such considered were taken to be either second order stationary, see Sec. 5.3.1, or described by second order stationary increments, see Sec. 5.3.2. We showed that analytic solutions exist for certain cases where the price and the return processes are described by *i.i.d.* or by auto-regressive fluctuations. These analytic solutions were compared with numerical results for situations where auto-covariance matrices have to be estimated from finite samples. We found that in general, the sampling noise leads to an underestimation of risk, but that asymptotic results are well approximated when samples used to estimate auto-covariance matrices are sufficiently large.

Looking at S&P500 data, we found that (normalised) auto-covariance spectra closely resemble those one would expect for price processes with independent increments, see Fig. 5.6, and it is this observation that motivates our choice of target matrix within a shrinkage cleaning strategy. To restate, shrinkage cleaning is the statistical estimation methodology where one takes a convex combination of a target and empirical estimator – and finds the best combination of the two that reduces the noisiness of estimation. We observed that the auto-covariance matrix cleaning gives rise to smoother trading strategies, and that it also leads to a reduction of risk in risk-return profiles, shown in Fig. 5.7.

A generalisation of the present work would expand the model to deal with a multi-period multi-asset version of a mean-variance formulation of optimal trading strategies. We are not aware of an investigation of the effects of sampling noise in the multi-period multi-asset case. Indeed the spectral theory of that case which would be useful to motivate and design cleaning strategies has not been developed as of now. Another direction that could be pursued is to include higher moments of strategy-return distributions in measures of risk, to better capture risk in the presence of fat-tailed return distributions. The translation into the time-domain, as advocated in the present study would, in general, involve k-point correlations of returns in time (where $k \geq 3$). Assessing sampling noise in such a situation would then clearly transcend the realm of random matrix theory.

To summarise, our goal was to explore two different forms of price formation in financial and gambling markets. For gambling we found the behaviour of odds signals to be similar – but not equivalent – to financial price. The initial odds in the horse racing in-play markets were found to behave in the same manner as the division of a randomly cut interval [4]. Toy models of the in-play odds signals were constructed along with statistical arbitrage trading strategies. We took the BHM model and created an agent-based trading model [129]. The BHM was also generalised such that the assumption of a linear information rate function was expanded to handle non-linear functions and both of which were fitted to winning horse in-play odds signals. The final part of the thesis was only concerned with financial price and took the Markowitz portfolio model and translated it into the time domain – giving optimal trading strategies as presented in [6].

References

[1] J. W. Tukey. *Exploratory Data Analysis (Behavioral Science)*. Pearson, 1977. vii, 28, 29

[2] L. V. Williams and D. S. Siegel, editors. *The Oxford Handbook of the Economics of Gambling*. Oxford, 2014. viii, 9, 42

[3] Sam Priestley and Toby Aldous. *P.A.S. CAPITAL LIMITED*. 2014. 1, 3, 5, 20, 180

[4] P. A. Bebbington and J. Bonart. Order statistics of horse racing gambling and random divisions of an interval, 2016. (Working Paper). 3, 34, 41, 185

[5] H. M. Markowitz. Portfolio selection. *The Journal of Finance*, 7:77–91, 1952. 6

[6] P. A. Bebbington and R. Kühn. Optimal trading strategies–a time series approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2016. 6, 146, 186

[7] E. Aurell and P. Muratore-Ginanneschi. Growth-optimal strategies with quadratic friction over finite-time investment horizons. *International Journal of Theoretical and Applied Finance*, 7:645–657, 2004. 6

[8] K. Muthuraman and S. Kumar. Multidimensional portfolio optimization with portional transaction costs. *Mathematical Finance*, 16:301–335, 2006. 6

[9] K. Muthuraman and H. Zha. Simulation-based portfolio optimization for large portfolios with transaction costs. *Mathematical Finance*, 18:115–134, 2008. 6

[10] A. W. Lynch and S. Tan. Multiple risky assets, transaction costs, and return predictability: Allocation rules and implications for us investors. *Journal of Financial and Quantitative Analysis*, 45:1015–1053, 2010. 6

[11] S. Basak and G. Chabakauri. Dynamic mean-variance asset allocation. *Review of Financial Studies*, 23:2970–3016, 2010. 6

[12] D. B. Brown and J. E. Smith. Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds. *Management Science*, 57:1752–1770, 2011. 6

[13] N. B. Garleanu and L. H. Pedersen. Dynamic trading with predictable returns and transaction costs. *Journal of Finance*, 68:2309–2340, 2013. 6

[14] V. DeMiguel, X. Mei, and F. J. Nogales. Multiperiod portfolio optimization with many risky assets and general transaction costs. Available at SSRN: http://ssrn.com/abstract=2295345. 6

[15] G. Connor, L. R. Goldberg, and R. A. Korajczyk. *Portfolio Risk Analysis*. Princeton University Press, Princeton and Oxford, 2010. 6

[16] E. J. Elton, M. J. Gruber, S. J. Brown, and W. N. Goetzmann. *Modern Portfolio Theory and Investment Analysis*. Wiley, New York, 2010. 6

[17] K. E. Back. *Asset Pricing and Portfolio Choice Theory*. OUP, 2010. 6, 57, 58, 72, 73, 74

[18] J. P. Bouchaud and M. Potters. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, Cambridge, 2006. 6, 57, 58, 72, 73, 74, 138, 147

[19] A. Meucci. *Risk and Asset Allocation*. Springer, 2005. 6

[20] Betfair. `https://www.betfair.com`. 7, 8, 12, 20, 26

[21] H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley, 2003. 7, 35, 40

[22] S. M. Gainsbury. *Internet Gambling: Current Research Findings and Implications*. Springer-Verlag New York, 2012. 7, 9, 11

[23] H2 Gambling Capital mobile gaming and betting gross win to reach 45% by 2018, new report by h2 gambling capital reveals retrieved december 04, 2014, from. http://www.igamingbusiness.com/news/mobile-gaming-and-betting-gross-win-reach-45-2018-new-report-\h2-gambling-capital-reveals. 7

[24] J. E. Granta, B. L. Odlaugb, and S. R. Chamberlainc. Neural and psychological underpinnings of gambling disorder: A review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 65:188–193, 2016. 8, 11

[25] S. M. Gainsbury, A. Russell, A. Blaszczynski, and N. Hing. The interaction between gambling activities and modes of access: A comparison of internet-only, land-based only, and mixed-mode gamblers. *Addictive Behaviors*, 41:34–40, February 2015. 8

[26] P. Gomber, P. Rohr, and U. Schweickert. Sports betting as a new asset class—current market organization and options for development. *Financial Markets and Portfolio Management*, 22(2):169–192, 2008. 9

[27] M. A. Smith & L. V. Williams. *Betting Exchanges: A Technological Revolution in Sports Betting*, chapter 19, pages 403–418. Handbook of Sports and Lottery Markets, 2008. 9, 15

[28] J. M. Keynes. *The General Theory of Employment, Interest and Money*. Palgrave Macmillan, 1935. 9

[29] D. B. Hausch and W. T. Ziemba, editors. *Handbook of Sports and Lottery Markets*. North Holland, first edition, October 2008. 9, 14

[30] T. Piketty. *The Economics of Inequality*. Harvard University Press, 2015. 10

[31] T. Piketty. About capital in the twenty-first century. *American Economic Review: Papers & Proceedings*, 105(5):48–53, 2015. 10

[32] L. H. Pedersen. *Efficiently Inefficient: How Smart Money Invests and Market Prices Are Determined*. Princeton University Press, 2015. 10

[33] European Gaming and Betting Association. Written submission to the green paper on online gambling in the internal market. Brussels: European Gaming and Betting Association., 2011. 11

[34] M. M. Ali. Some evidence of the efficiency of a speculative market. *Econometrica*, 47(2):387–392, March 1979. 14

[35] Y. Hirono and Y. Hidaka. Jarzynski-type equalities in gambling: Role of information in capital growth. *Journal of Statistical Physics*, 161:721–742, 2015. 14

[36] A. Ozgit. Posted-offer vs. double auctions revisited: An investigation into online sports betting, 2005. 15

[37] J. R. Lowery P. J. Healy, S. Linardi and J. O. Ledyard. Prediction markets: Alternative mechanisms for complex environments with few traders. *Management Sci*, 56(11):1977–1996, 2010. 15

[38] J. Haigh and L. V. Williams. *Index Betting for Sports and Stock Indices,*, chapter 17, pages 357–383. Elsevier B.V., 2008. 15

[39] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. 29

[40] G. Corrado. Measurement of inequality of incomes. *The Economic Journal*, 31(121):124–126, 1921. 31

[41] P. Artzner, F. Delbaen, J. M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999. 31

[42] H. Theil. *Economics and Information Theory.* Amsterdam: North-Holland, 1967. 31

[43] J. v. Neumann. *Mathematical Foundations of Quantum Mechanics.* Springer, 1932. 31

[44] A. B. Atkinson. On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263, September 1970. 31

[45] A. F. Shorrocks. The class of additively decomposable inequality measures. *Econometrica*, 48(3):613–625, April 1980. 32

[46] L. Holst. On the length of the pieces of a stick broken at random. *Journal of Applied Probability*, 17(3):623–634, 1980. 34, 40

[47] R. J. Elliott, J. V. D. Hoek, and W. P. Malcolm. Pairs trading. *Quanitive Finance*, 2005. 46, 147

[48] B. Do, R. Faffans, and K. Hamza. A new approach to modeling and estimation for pairs trading. *Working Paper, Monash University.*, 2006. 46, 147

[49] B. Do and R. Faff. Does simple pairs trading still work? *Financial Analysts Journal*, 66(4), 2010. 46, 147

[50] D. Herlemont. Pairs trading, convergence trading cointegration, 2013. 46, 147

[51] R. Q. Liew and Y. Wu. Pairs trading: A copula approach. *Journal of Derivatives & Hedge Funds*, 19(1):12–30, 2013. 46, 147

[52] C. Krauss. Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 2016. 46, 147

[53] F. A. Hayek. The use of knowledge in society. *The American Economic Review*, 35(4):519–530, 1945. 47

[54] G. R. Gibson. *The stock exchanges of London, Paris, and New York: a comparison*. New York; London: G.P. Putnam, 1889. 47

[55] J. Regnault. *Calcul des chances et philosophie de la bourse*. Mallet-Bachelier et Castel, Paris, 1863. 47

[56] L. Bachelier. *Théorie de la spéculation*. PhD thesis, University of Paris, 1900. 47

[57] P. A. Samuelson. Proof that properly anticipated prices flucuate randomly. *Industrial Management Review*, 6(2):41–49, 1965. 47, 48

[58] A. Einstein. On the movement of small particles suspended in stationary liquids required by the molecular-kinetic theory of heat. *Annalen Der Physik*, 1905. 47

[59] E. F. Fama. Random walks in stock market prices. *Financial Analysts Journal*, 21(5):55–59, 1965a. 48

[60] E. F. Fama. The behavior of stock-market prices. *Journal of Business*, 38(1):34–105, 1965b. 48

[61] E. F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970. 48

[62] M. Sewell. History of the efficient market hypothesis. Online, Jan 2011. 48

[63] S. Nadarajah and S. Kotz. Exact distribution of the max/min of two gaussian random variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2008. 52

[64] W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442., 1964. 54

[65] M. Capiński and E. Kopp. *Measure, Integral and Probability.* Springer-Verlag, New York, 2004. 57, 58, 72, 73, 74

[66] J. Jacod and P. Protter. *Probability Essentials.* Springer-Verlag, Berlin, 2004. 57, 58, 72, 73, 74

[67] M. Musiela and M. Ruthowski. *Martingale Methods in Financial Modelling.* Springer-Verlag, Berlin, 2005. 57, 58, 72, 73, 74

[68] S. Shreve. *Stochastic Calculus for Finance; Volume II: Continuous-Time Models.* Springer-Verlag, New York, 2004. 57, 58, 72, 73, 74, 147

[69] S. Shreve. *Stochastic Calculus for Finance; Volume I: The Binomial Asset Pricing Models.* Springer-Verlag, New York, 2005. 57, 58, 72, 73, 74

[70] D. Brigo and F. Mercurio. *Interest Rate Models - Theory and Practice: With Smile, Inflation and Credit*. Springer-Verlag, Berlin, 2007. 57, 58, 72, 73, 74

[71] C. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer-Verlag, Berlin, 2009. 57, 58, 72, 73, 74

[72] A. Pascucci. *PDE and Martingale Methods in Option Pricing*. Springer-Verlag, Berlin, 2011. 57, 58, 72, 73, 74

[73] G. E. Uhlenbeck and L. S. Ornstein. On the theory of brownian motion. *Physical Review Letters*, 36:823–841, 1930. 57

[74] D. A. Dickey and W. A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:115–143, 1979. 58

[75] D. Kwiatkowski, P. C. B. Philips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54:159–178, 1992. 58, 150

[76] H. E. Hurst. Long-term storage capacity of reservoirs. *Transactions of American Society of Civil Engineers*, 116(1):770–799, 1951. 58, 147

[77] J. Kalda. Oceanic coastline and super-universality of percolation clusters. *ARXIV*, 2002. 59

[78] I. Chattoraj, S. Tarafder, and M. Nasipuri. Fractal analysis to determine self-similar characteristics in the microstructure of hsla steel. *Materials and Manufacturing Processes*, 24(2):145–149, 2009. 59

[79] M. Tarafder, I. Chattoraj, S. Tarafder, and M. Nasipuri. Self-similar and self-affine characteristics of microstructural images of hsla steel. *Materials Science and Technology*, 25(4):542–548, 2009. 59

[80] A. Charles and O. Darné. Variance ratio tests of random walk: An overview. *Journal of Economic Surveys*, 23(3):503–527, 2009. 60

[81] J. D. Hamilton. *Time Series Analysis.* Princeton University Press, Princeton and Oxford, 1994. 61, 138, 147, 150

[82] W. F. Sharpe. Mutual fund performance. *Journal of Business*, 39(S1):119–138, 1966. 66

[83] A. Macrina. An Information-Based Framework for Asset Pricing: X-factor Theory and its Applications. Technical report, KCL, 2006. 69, 72, 73, 74, 76, 78, 81

[84] D. C. Brody, L. P. Hughston, and A Macrina. *Beyond Hazard Rates: a New Approach to Credit Risk Modelling*, pages 231–257. Birkhauser, 2007. 69, 72, 73, 74, 76, 78

[85] D. C. Brody, L. P. Hughston, and A. Macrina. Dam rain and cumulative gain. *Proceedings of the Royal Society London A*, 464(2095):1801–1822, 2007. 69, 72, 73, 74, 76, 78

[86] D. C. Brody, L. P. Hughston, and A. Macrina. Information-based asset pricing. *IJTAF*, 11:107–142, 2008. 69, 74, 76, 78, 81, 86, 143, 182

[87] D. C. Brody, L. P. Hughston, and A. Macrina. Information-based asset pricing. *International Journal of Theoretical and Applied Finance*, 11(1):107–142, 2008. 69, 74, 76, 78, 143, 182

[88] L. P. Hughston and A. Macrina. Information, inflation, and interest. In L. Stettner, editor, *Advances in Mathematics of Finance*, volume 83, pages 117–138. Banach Center Publications, Polish Academy of Sciences, 2008. 69, 74, 76, 78, 143, 182

[89] A. Capponi and J. Cvitanić. Credit risk modeling with misreporting and incomplete information. *International Journal of Theoretical and Applied Finance*, 12(01):83–112, 2009. 69, 74, 76, 78, 143, 182

[90] D. C. Brody, M. H. A. Davis, R. L. Friedman, and L. P. Hughston. Informed traders. *Proceedings of the Royal Society A Ñ Mathematical, Physical & Engineering Sciences*, 465:1103–1122, 2009. 69, 74, 76, 78, 92, 143, 182

[91] D. C. Brody, L. P. Hughston, and A. Macrina. *Credit risk, market sentiment and randomly-timed default*, pages 267–280. Springer Verlag, 2010. 69, 74, 76, 78, 143, 182

[92] D. C. Brody and Y. T. Law. Theory of information pricing, April 2010. 69, 74, 76, 78, 143, 182

[93] A. Macrina and P. A. Parbhoo. *Security Pricing with Information-Sensitive Discounting*, pages 157–180. World Scientific Publishing Company, 2010. 69, 74, 76, 78, 143, 182

[94] D. C. Brody, L. P. Hughston, and A. Macrina. *Modelling Information Flows in Financial Markets*, chapter 5, pages 133–153. Springer-Verlag, 2011. 69, 74, 76, 78, 94, 143, 182

[95] E. Hoyle, L. P. Hughston, and A. Macrina. Lévy random bridges and the modelling of financial information. *Stochastic Processes and their Applications*, 121(4):856–884, 2011. 69, 74, 76, 78, 143, 182

[96] J. Akahori and A. Macrina. Heat kernel interest rate models with time-inhomogeneous markov processes. *International Journal of Theoretical and Applied Finance*, 15(1), 2012. 69, 74, 76, 78, 143, 182

[97] P. V. Gapeev. Pricing of perpetual american option in a model with partial information. *International Journal of Theoretical & Applied Finance*, 15(1):1250010–1 – 1250010–21, 2012. 69, 74, 76, 78, 143, 182

[98] D. Filipovi, L. P. Hughston, and A. Macrina. Conditional density models for asset pricing. *International Journal of Theoretical and Applied Finance*, 15(1):1–24, 2012. 69, 74, 76, 78, 143, 182

[99] D. Filipovic, L. P. Hughston, and A. Macrina. Conditional density models for asset pricing. *International Journal of Theoretical and Applied Finance*, 15(1):1–24, 2012. 69, 74, 76, 78, 143, 182

[100] L. P. Hughston and A. Macrina. Pricing fixed-income securities in an information-based framework. *Applied Mathematical Finance*, 19(4):361–379, 2012. 69, 74, 76, 78, 143, 182

[101] D. C. Brody, B. K. Meister, and M. F. Parry. Informational inefficiency in financial markets. *Mathematics and Financial Economics*, 6(3):249–259, 2012. 69, 74, 76, 78, 143, 182

[102] D. C. Brody, L. P. Hughston, and X. Yang. Signal processing with lévy information. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 469(2149), 2012. 69, 74, 76, 78, 143, 182

[103] D. C. Brody, L. P. Hughston, and E. Mackie. General theory of geometric lévy models for dynamic asset pricing. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 468(2142):1778–1798, 2012. 69, 74, 76, 78, 143, 182

[104] D. C. Brody and L. P. Hughston. Lévy information and the aggregation of risk aversion. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 469(2154), April 2013. 69, 74, 76, 78, 143, 182

[105] U. Horst, M. Kupper, A. Macrina, and C. Mainberger. Continuous equilibrium in affine and information-based capital asset pricing models. *Annals of Finance*, 9(4):725–755, 2013. 69, 74, 76, 78, 143, 182

[106] E. Hoyle, L. P. Hughston, and A. Macrina. Stable-1/2 bridges and insurance. In A. Palczewski and L. Stettner, editors, *Advances in Mathematics of Finance*. Banach Center Publications, Polish Academy of Science, Institute of Mathematics, 2014. 69, 74, 76, 78, 143, 182

[107] A. Macrina. Heat kernal models for asset pricing. *International Journal of Theoretical and Applied Finance*, 17(7):1450048, 2014. 69, 74, 76, 78, 143, 182

[108] A. Macrina and J. Sekine. Filtering with randomised markov bridges. 2014. 69, 74, 76, 78, 143, 182

[109] A. Macrina and P. A. Parbhoo. Randomised mixture models for pricing kernels. *Asia-Pacific Financial Markets*, 21(4):281–315, 2014. 69, 74, 76, 78, 143, 182

[110] D. C. Brody and Y. T. Law. Pricing of defaultable bonds with random information flow. *Applied Mathematical Finance*, 22(5):399–420, 2015. 69, 74, 76, 78, 143, 182

[111] N. Arora, J. R. Bohn, and F. Zhu. Reduced form vs. structural models of credit risk: A case study of three models. *Journal of Investment Management*, 3(4), 2005. 69

[112] M. Ammann. *Credit Risk Valuation*. Springer, 2002. 69

[113] R. Merton. On the Pricing of Corporate Debt: The Risk Structure of Interest Rates. *Journal of Finance*, 29(2):449–470, 1974. 70

[114] O. A. Vasicek. Credit valuation. *KMV*, March 1984. 70

[115] Y. Wang. Structural credit risk modeling: Merton and beyond. *Risk Management*, 2009. 70

[116] R. Jarrow and S. Turnbull. Credit risk: drawing the analogy. *Risk Magazine*, 5(9), 1992. 70

[117] R. Jarrow and S. Turnbull. Pricing derivatives on financial securities subject to credit risk. *Journal of Finance*, 50(1), 1995. 70

[118] J. C. Hull and A. D. White. Valuing Credit Default Swaps I No Counterparty Default Risk. *The Journal of Derivatives*, 8(1):29–40, 2000. 70

[119] U. Cetin, R. Jarrow, P. Protter, and Y. Yildrim. Modelling credit risk with partial information. *Annals of Applied Probability*, 14:1167–1172, 2004. 70

[120] R. A. Jarrow and P Protter. Structural versus reduced form models: A new information based perspective. *The Journal of Investment Management*, 2(2):1–10, 2004. 70

[121] I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus.* Springer-Verlag, second edition, 1991. 75

[122] J. P. Bouchaud, M. Mezard, and M. Potters. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2(4):251–256, 2002. 79

[123] E. Smith, J. D. Farmer, L. Gillemot, and S. Krishnamurthy. Statistical Theory of The Continuous Double Auction. *Quant. Finance*, 3:481–514, 2003. 79

[124] J. D. Farmer, L. Gillemot, G. Iori, S. Krishnamurthy, D. E. Smith, and Marcus G. Daniels. A Random Order Placement Model of Price Formation in the Continuous Double Auction. *The Economy as an Evolving Complex System*, 3:133–173, 2005. 79

[125] R. Cont, A. Kukanov, and S. Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 0(0):1–42, 2013. 79, 182

[126] D. Williams. *Probability with Martingales.* Cambridge University Press, 1991. 83

[127] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), July 1948. 91

[128] M. Mézard and A. Montanari. *Information, Physics, and Computation.* Oxford University Press, 2009. 91, 92

[129] P. A. Bebbington, I. J. Ford, and F. M. C. Witte. Information based trading. (Working Paper), 2017. 93, 185

[130] L. C. G. Rogers & D. Williams. *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations.* Cambridge University Press, second edition, 2000. 126

[131] L. C. G. Rogers & D. Williams. *Diffusions, Markov Processes and Martingales: Volume 2, Itô Calculus.* Cambridge University Press, second edition, 2000. 126

[132] H. M. Mahmoud. *Sorting: A Distribution Theory.* John Wiley & Sons Inc, 2011. 128

[133] L. Laloux, P. Cizeau, J. P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, 83:1467–1470, 1999. 145, 146, 175

[134] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley. Universal and non-universal properties of cross-correlations in financial time series. *Phys. Rev. Lett.*, 83:1471–1474, 1999. 145

[135] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization ii. *Physica A: Statistical Mechanics and its Applications*, 319:487–494, March 2003. 145, 146

[136] O. Ledoit and M. Wolf. Honey, I shrunk the sample covariance matrix. *J. Portfolio Management*, 31:110–119, 2004. 145, 146, 175

[137] G. Papp, S. Pafka, M. A. Nowak, and I. Kondor. Random matrix filtering in portfolio optimization. *Acta Physica Polonica B*, 36(9):2757–2765, 2005. 145, 146

[138] L. Laloux. M. Potters, J. P. Bouchaud. Financial application of random matrix theory: Old laces and new pieces. *Acta Phys. Pol. B*, 36:2767, 2005. 145, 146

[139] V. Golosnoy and Y. Okhrin. Multivariate shrinkage for optimal portfolio weights. *The European Journal of Finance,*, 13:441–45, 2007. 145, 146

[140] Y. Chen, A. Wiesel, and A. O. Hero. Shrinkage estimation of high dimensional covariance matrices. In *IEEE Intl Conf. on Acoust., Speech, and Signal Processing*, 2009. 145, 146

[141] S. Still and I. Kondor. Regularizing portfolio optimization. *New Journal of Physics*, 12:075034, 2010. 145, 146

[142] O. Ledoit and M. Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40:1024–1060, 2012. 145, 146, 175

[143] F. Caccioli, S. Still, M. Marsili, and I. Kondor. Optimal liquidation strategies regularize portfolio selection. *The European Journal of Finance*, 19:554–571, 2013. 145, 146

[144] F. Caccioli, I. Kondor, and M. Marsili amd S. Still. l_p regularized portfolio optimization. http://arxiv.org/pdf/1404.4040.pdf, April 2014. 145, 146

[145] J. Wishart. Generalized Product Moment Distribution in Samples. *Biometrika*, 20 A:32–52, 1928. 145

[146] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb.*, 1:457–483, 1967. 145, 146

[147] D. Li and W. L. Ng. Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Mathematical Finance*, 10:387–406, 2000. 146, 177

[148] R. Kühn and P. Sollich. Spectra of empirical auto-covariance matrices. *EPL (Europhysics Letters)*, 99:20008, 2012. 146, 151, 158

[149] K. P. Murphy. *Machine Learning: a Probabilistic Perspective.* MIT Press, 2012. 147

[150] B. Gerrard. *The role of econometrics in a radical methodology*, volume 1, chapter 8, pages 110–132. Edward Elgar, 2002. 147

[151] M. Pirooznia, S. R. Emadi, and M. N. Alamdari. The time series spectral analysis of satellite altimetry and coastal tide gauges and tide modeling in the coast of caspian sea. *Open Journal of Marine Science*, 6(2), 2016. 147

[152] G. Gudmundsson. Time-series analysis of imports, exports and other economic variables. *Journal of the Royal Statistical Society. Series A (General)*, 134(3), 1971. 147

[153] L. D. Persioa. Autoregressive approaches to importĐexport time series i: basic techniques. *Modern Stochastics: Theory and Applications*, 2(1):51–65, 2015. 147

[154] H. J. Bierens. Information criteria and model selection, 2006. 147

[155] P. J. Brockwell. *Introduction to Time Series and Forecasting*. Springer, 2002. 147

[156] R. S. Tsay. *Analysis of Financial Time Series*. Wiley, third edition, 2010. 147

[157] D. Barber, A. T. Cemgil, and S. Chiappa. *Bayesian Time Series Models*. Cambridge University Press, 2015. 147

[158] W. K. Newey and K. D West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. 150

[159] J. P. Bouchaud and M. Potters. Financial applications of random matrix theory: a short review. In G. Akemann, J. Baik, and P. D. Francesco, editors, *The Oxford Handbook of Random Matrix Theory*. Oxford University Press, Oxford, 2011. 174, 175

[160] C. Stein. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution. *Math. Rev.*, 1:197–206, 1956. 175

[161] M. Tumminello, F. Lillo, and R. N. Mantegna. Shrinkage and Spectral Filtering of Correlation Matrices: a Comparison Via the Kullback-Leibler Distance. *Act. Phys. Pol.*, B 38:4079–4088, 2007. 175

[162] J. P. Bouchaud. Price impact, 2009. 182

[163] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, first edition, 2015. 184