Nonlinearity in high-frequency finance and optimal execution

Gianbiagio Curato

A Dissertation
Presented to
Scuola Normale Superiore
in Candidacy for the Degree
of *Philosophiae Doctor* in Mathematics for finance

Supervisor: Prof. Fabrizio Lillo

Academic Year: 2014/2015

Abstract

In this thesis we consider two problems regarding the field of market microstructure: the analysis and modeling of high frequency financial data and the optimal execution of large orders. The first problem is analyzed empirically within the context of the dynamics of limit order books. Such context imposes a discrete modeling approach about the variables of interest, like prices and bid-ask spreads. The optimal execution problem, instead, is analyzed within the context of a price model which allows for a nonlinear-transient market impact [43, 104]. Our interest is motivated by the fact that the study and implementation of optimal strategies is one of the main tasks of large banks and capital investment firms. The two problems are defined by two different nonlinear models. A nonlinear modeling approach is necessary in order to reproduce the empirical data. Our work focuses on the analysis of the implications of such nonlinearities on the stochastic dynamics of prices and on optimal strategies.

In the first part of the thesis, the main idea behind our modeling approach is that the description of price changes, recorded at time scales of the single transactions, should be described by discrete variables in the context of a nonlinear econometric model. This implies a different modeling approach for prices with respect to stochastic processes like the arithmetic or geometric Brownian motions. The different modeling is required particularly in order to reproduce the empirical properties of a specific class of stocks, i.e. large tick[1] stocks. This kind of stocks are described by a relatively high tick size/price ratio and by a bid-ask spread that is almost always one. The interest in the study of such kind of stocks is also motivated by the current debate on the optimal tick size in order to design financial markets. Even though we do not address directly the problem of the optimal tick size, our study highlights the fact that for the class of large tick stocks the price dynamics is heavily affected by the discrete nature of the price grid. This analysis is developed in chapters 3, 4, 5.

The result of this first part of the thesis work is the design of a model for high frequency price dynamics, whose time evolution is defined in transaction time[2], that is able to reproduce the empirical observations regarding price changes and bid-ask spread processes for the class of large tick stocks. We focus on the effort to reproduce as many empirical observations as possible for what regards the price change process. The key observation, highlighted in chapter 3, is that the price changes distributions recorded from time scales greater than few transactions have different populations between even and odd price changes. In contrast to this observation, the distribution of price changes recorded at the time scale of one transaction is not described by different populations between even and odd values, i.e. the only relevant observation is that at this time scale the price changes are almost always zero. In order to

[1]The tick size is the smallest increment by which the price of stocks, futures contracts or other exchange-traded instrument can move.

[2]The transaction time is a discrete time measure defined by an integer value that increases by one unit when a transaction, or a group of transactions recorded at the same time in the database of interest, is executed.

reconcile these two contrasting empirical observations, we propose to couple the dynamics of price changes to the dynamics of the bid-ask spread. This coupling is modeled by a Markov switching process in chapter 4. The coupling enable us to reproduce the empirical price changes distributions at all time scales under exam, i.e. ranging from the scale of a single transaction to the daily time scale. How well such distributions are reproduced is the main topic of the chapter 5. The results of the chapter 5 confirm the importance of the coupling in order to reproduce price changes distributions. Another empirical observation leads to an improvement of the Markov switching modeling approach. We observe high persistent correlations between the squares of price changes at the time scale of single transactions. In order to reproduce the persistent correlations between the square of price changes, that is a proxy of the high frequency volatility of the stock, we need a model more advanced of the Markov switching process. We introduce in chapter 4 a double chain Markov process (DCMM) in order to model simultaneously the coupling between bid-ask spread and price changes, and the persistent correlations between the square of price changes. Such process shows to be able to model the persistent correlations of squared price changes observed empirically.

In the second part of our work we address the problem of finding the optimal execution strategy for a price model characterized by a nonlinear-transient market impact, also known as the propagator model. The optimal strategy is usually found by the minimization of a certain cost functional defined by the price model under exam. We study optimal strategies obtained as minimizers of the implementation shortfall cost function in a risk-neutral case. The existence, or not, and the specific structure of optimal strategies are fundamental properties in order to study possible price manipulations allowed by the specific model under exam. It is clear that a well-behaved price model should not allow for price manipulations. This statement is similar to characterizing the absence of arbitrage in an asset pricing model. The particular model and cost function under exam allow to search for statically optimal strategies that are also dynamically optimal. This reduces the problem to the study of a deterministic trading strategy, which is the global minimum of the cost functional. If one imposes the constraint that the trading rate never vanishes, it is possible to show that any local minimum of the cost function is described by a nonlinear integral equation of the first kind of the Urysohn type [81]. We search for an approximated solution of such integral equation, developing an original semi-analytic approach based on the homotopy technique. In order to search the global minimum, instead, we consider a direct derivative-based numerical optimization of the cost function. At the end of this procedure, we compare the costs of the obtained optimal strategies from the two methods, showing that the global minimum has cost much smaller than the one obtained with the integral equation. The strategy, describing the global minimum, alternates buys and sells and the corresponding cost is sometimes even negative, indicating the possibility of price manipulations. It is worth mentioning that imposing that one can only buy during a buy execution leads to corner solutions. These corresponds to strategies where it is optimal not to trade in some subintervals.

Since the derivative of the cost function diverges for zero trading rates, we use in this case derivative-free optimization methods.

The main result of the second part of the thesis work regards the analysis of the shape and expected costs of the optimal execution strategy, i.e. a function describing the trading rate with respect to the real time, and the specific kind of price manipulations allowed by the propagator model. Moreover, the explicit form of the equations computed by the homotopy analysis method for the nonlinear integral equation is an original application of the homotopy analysis to a financial case of study. The homotopic solution for the optimal strategy, developed in chapter 7, is affected by a specific coupling between the present and future values of the trading rate caused by the nonlinearity of the market impact function. This coupling disappears in the linear impact case. The presence of such coupling imposes to use the homotopy analysis technique valid for a bi-dimensional system. The final approximated homotopic solution is different with respect to the linear impact case. For example, for a buy program one has to buy faster in the first half of the trading period. The solution valid in the linear impact case, instead, is symmetric respect to the mid of the trading period [108]. Moreover, the homotopic solution has a cost lower than classic strategies, like for example a VWAP strategy.

In the case of a concave market impact function, the integral equation describes only a local minimum of the cost function. In order to search for a global minimum, in chapter 8 the optimal strategy is analyzed by a direct minimization of the cost functional. This numerical analysis shows that the propagator model allows for two possible kinds of price manipulations, i.e. transaction triggered price manipulations and negative expected cost strategies. We propose to regularize the propagator model by mean of two distinct terms, i.e the first is a bid-ask spread cost and the second is a concave-convex price impact function. Both of them are able to rule out the presence of negative expected cost strategies. The presence of transaction triggered price manipulations is, instead, only mitigated by the proposed regularizing terms.

In chapter 1 we introduce the main topic of the thesis: the importance of nonlinearities in modeling and characterization of high frequency financial data and optimal execution strategies.

In chapter 2, we review some properties of electronic markets and limit order books and the main quantities that define them. We explain in some detail the notion of tick size, because it is important for the development of the model reported in chapter 4.

In chapter 3, we perform an extensive empirical analysis of the impact of the tick size, i.e. the smallest allowed difference between prices, on the distributional properties of the bid-ask spreads, price changes and returns for NASDAQ stocks. Some new empirical observations are presented as motivation for the development of the model presented in the chapter 4.

In chapter 4, we present a new model for the coupled dynamics of bid-ask spreads and price-changes in the context of Markov-switching models. In order to reproduce the persistent correlations between squared price changes we introduce a double chain Markov model (DCMM). This model is able to reproduce the empirical properties of the price changes dynamics of large tick stocks traded at NASDAQ market.

In chapter 5, we propose a multiscale model selection method based on the Jensen-Shannon distance in order to select between different models for price changes. The Jensen-Shannon distance shows that the coupling between spreads and price changes is important to model price changes distributions at different time scales of observations.

In chapter 6, we review the properties of market impact of trades on prices and the problem of optimal execution of large orders. We are mainly interested to show the different kinds of price manipulations that can affect in general a price impact model. We summarize the constraints that the absence of such manipulations imposes to the Almgren-Chriss model [16] and to the propagator model [43, 104].

In chapter 7, we study optimal trading strategies in presence of a nonlinear market impact in the context of the propagator model. We develop and implement an original semi-analytic procedure, designed using the homotopy analysis method, in order to solve the nonlinear Urysohn equation defining the local minimum of the cost functional.

In chapter 8, we adopt a fully numerical approach to find the global minimum of the expected cost functional. We make use of two different kinds of algorithms, i.e the sequential quadratic programming algorithm (SQP) and direct search methods. This analysis shows that the presence of a concave market impact implies a cost functional described by a rugged landscape, i.e there are many well-separated minima. For a buy strategy the structure of each minimum is described by few bursts of buying interspersed by periods of weak selling. Moreover, such burst strategies have a cost much lower than that relative to the homotopic strategies.

In appendix A we perform a study on a specific deterministic neural network model. The motivation of this study is due to our interest in the study of dynamical systems. The neural model is defined by a nonlinear coupling between neurons. The main result is the characterization of the chaotic attractor of the neural network by means of Lyapunov exponents.

Keywords: tick size, large tick stocks, Markov switching processes, Double chain Markov processes, bid-ask spread, price changes, volatility clustering, Jensen-Shannon distance, nonlinear transient market impact, propagator model, optimal order execution, Urysohn equation of the first kind, SQP algorithm, direct search algorithm, bursts strategies, neural networks, Lyapunov exponents.

List of publications

1. Curato, G., & Politi, A. (2013). Onset of chaotic dynamics in neural networks. *Physical Review E*, 88(4), 042908.

2. Curato G, Lillo F. Multiscale Model Selection for High-Frequency Financial Data of a Large Tick Stock by Means of the Jensen-Shannon Metric. *Entropy*. 2014; **16**(1):567-581.

3. Curato, G., Gatheral, J., & Lillo, F. (2014). Optimal execution with nonlinear transient market impact. Available at *SSRN* 2539240.

4. Curato, G., & Lillo, F. (2015). How Tick Size Affects the High Frequency Scaling of Stock Return Distributions. In *Financial Econometrics and Empirical Market Microstructure* (pp. 55-76). Springer International Publishing.

5. Curato, G., & Lillo, F. (2015). Modeling the coupled return-spread high frequency dynamics of large tick assets. *J. Stat. Mech.* 2015 P01028 doi:10.1088/1742-5468/2015/01/P01028.

Chapters 1, 2 and 6 have an introductory purpose and contain mainly non original work. Chapter 3 covers some of the contents of the article number 4. Chapter 4 covers the content of the article number 5 and contains some extensions not covered in the article. Chapter 5 covers the content of the article number 2. Chapters 7 and 8 cover the content of the article number 3. The appendix A covers the work developed in the article number 1.

Acknowledgements

I am truly indebted with my advisor professor Fabrizio Lillo, for encouraging me to spend the last years studying beautiful and challenging problems about mathematical finance. I have been consistently borrowing his advices and profiting of his skills, which he never refused to share. I thank professor Jim Gatheral, with whom i had the pleasure to collaborate with during the course of the PhD on the subject of optimal order execution. I thank professor Antonio Politi, with whom i had the pleasure to collaborate with on the subjects of neural networks dynamics and CUDA parallel programming paradigm. I acknowledge Alessandro Profeti and Andrea Carlo Giuseppe Mennucci for all the interesting discussions we had and for the exceedingly long time we spent about our common interests about computer science.

Finally, to my family and friends. To Mom, Dad and to my Sister, thank for your constant support and for your example of wisdom. To Valeria, Carlo, Sergio, Biru and Alessandra, thank for your help during a difficult period of my life.

This work was funded in part by the Research Grant: "Dinamiche di altissima frequenza, price impact e microstruttura dei mercati finanziari "at Scuola Normale Superiore.

To my parents.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is composed of two parts. The first part, developed in chapters 2, 3, 4, 5, deals with the analysis of a family of high frequency models for prices and bid-ask spreads. The models are designed in order to reproduce statistical properties of NASDAQ stocks. In the second part, developed in chapters 6, 7, 8, we study theoretically and numerically the problem of optimal order execution in the context of the propagator model [43, 104]. The common feature of these two models, i.e our model for high frequency price dynamics and the propagator model, is a nonlinear specification for the dynamics of the price process. In the first case, the nonlinear modeling approach succeeds in reproducing the empirical statistical properties of price and bid-ask spread dynamics. In the second case, a concave market impact implies the presence of optimal strategies with structures and properties much more complex than the one of the linear market impact case [108].

1.1 Modeling financial prices at high frequency

A famous climber, when asked why he was willing to put his life in danger to climb dangerous summits, answered: "Because they are there.". We would be tempted to give the same answer when people ask why we take so much pain in dealing with high-frequency financial data. The reason is simple: financial markets are the source of high-frequency data. The information regarding prices and the number of available shares, i.e the volume, in a modern electronic market is accessible at various level of detail. The most comprehensive kind of market data are Level 2 market data, i.e. they describe the entire depth of the market[1].

The meaning of the term *high-frequency* relies exactly on the fact that such data can be recorded at intradaily frequencies, i.e. ranging from hours to microseconds. By nature these data are irregularly spaced in time. Liquid markets generate hundreds of thousands of records, i.e. prices and volumes, per business day. Technological progress and the growing dominance of electronic trading allows to record market activity on high frequency and with high precision leading to advanced and comprehensive data sets. The introduction of electronic trading platforms have automatized and sped

[1]The market depth is the number of shares available for a given price value.

up trade execution as well as trading reporting and allow investors to automatize trading strategies, order routing, and real-time order management. Transaction and limit order book data, as provided by various exchanges and trading platforms, create strong academic interest as they allow to analyze the impact of institutional setting on the trading process, price discovery, and enable to study market dynamics and traders' behavior on the lowest possible aggregation level. Modeling limit order books on high frequencies provides insights into the interplay between liquidity supply and demand, execution risks, traders' order submission strategies, and the market impact of order placements. Addressing these issues requires the development and application of econometric models which are tailor-made for specific data and research tasks.

High-frequency data are becoming a fantastic experimental bench for understanding market microstructure and more generally for analyzing financial markets.

This leaves the researcher with the problem of dealing with such vast amounts of data using the right mathematical and statistical tools and models. Our approach in this thesis will follow three main steps. The first one is to explore the data in order to discover new fundamental statistical regularities of raw financial data. This is often called finding the *stylized facts* in the econometric or finance literature.

The second step is to use all of these empirical facts to formulate adequate models. By adequate models, we do not mean models that come from general hypothesis on financial markets, but rather models that are directly inspired by the empirical regularities encountered in the data. It is the point where our understanding of market behavior and reality of the data properties should meet. With few data, one tends to favor the simpler models because they contain few parameters and because test like the likelihood ratio test would strongly penalize the increase of parameters. Of course, simplicity is a desirable feature of theoretical models, but one should not seek simplicity at the cost of missing important features of the data-generating process. Sometimes, it is useful to explore more complicated models, i.e. which may contain more parameters. In the case of high-frequency data the penalty of using a large number of parameters is relatively small, because the abundance of the independently measured observations approximates an asymptotic environment. High-frequency data open also the way for studying financial markets at very different time scales, from milliseconds to years. This represents an available aggregation factor of several orders of magnitude. Some empirical properties are similar at different time scales, leading to fractal behaviors. Works on the scaling law of return volatility, bid-ask spreads, signs of trading orders, etc., have been flourishing in the past few years often coming from the physics community. It is a sign that the analysis of high-frequency financial data is moving toward a better understanding of the aggregation properties.

The third step, of course, is to verify whether these models satisfactorily reproduce the stylized facts found in the data. From a practical/industrial point of view, the data analysis and the corresponding modeling activity have as ultimate goal the ability to produce reasonable predictors of future price movements or risks and to integrate these tools into practical applications, such a risk management tool, option pricing algorithms or trading algorithms. The growing popularity of high-frequency finance is in fact triggered by technological progress in trading systems and trade recording as

well as an increasing importance of intraday trading, optimal trade execution, order placements and liquidity dynamics.

The classic approach used in the market microstructure literature [134] is a linear modeling describing continuous variables. The term *linear* here refers to all models that use ARMA or VAR processes for the description of the variables of interest. The VAR model of Hasbrouck [132], describing the mid quote and the sign of trades, the Roll model [217], and the model of Madhavan et al. (MRR) [177] fall into the category of continuous values linear models. Such models do not account for the intrinsic discreteness of financial high-frequency data, like prices, quotes, bid-ask spreads or trade counts in fixed intervals. Moreover, most high-frequency variables are positive valued, positively autocorrelated and strongly persistent, such kind of variables are for example the bid-ask spread and the available volume at the various price levels of the order book. This leads to a direct interest for modeling discrete valued data. In this thesis we develop a model for trade-to-trade price changes by means of discrete processes. The discrete modeling approach for prices is less used in literature than the continuous value approach, some examples of discrete price models are given by [221, 222, 163, 133]. One work on this direction of research was made by Russel and Engle [221]. They proposed a model where the distribution of each price change is a multinomial, conditional on past information and the time interval between the transactions.

A central topic of this thesis is the analysis of some *nonlinear models* for high frequency price dynamics. This implies the application of a nonlinear econometric model for price dynamics. Univariate nonlinear time series analysis has a long literature in econometric analysis. The most famous model is the autoregressive conditional heteroskedasticity model (ARCH) by Engle [93], and its generalized form (GARCH). Such models are stochastic volatility models, where the volatility is serially correlated. Some other popular nonlinear models such as bilinear models [119, 121], the threshold autoregressive model (TAR) [236], smooth transition autoregressive model (STAR) [58] or the Markov-switching autoregressive model (MSAR) [127], artificial neural networks models (ANN)[186, 203, 233], appeared only after the models had been developed by statisticians and applied to non-economic data. The number of economic applications has been growing steadily, and nonlinear models have been fitted to many macroeconomic and financial time series, both for forecasting and for testing implications of economic theories. The model developed in the first part of the thesis belongs to the class of Markov-switching models. In summary, we develop a nonlinear model for discrete price changes and discrete bid-ask spreads.

Other lines of research in economics have also tried to connect financial observations to the theory of deterministic nonlinear dynamics and chaos [213, 226, 169, 21]. Deterministic nonlinear processes are described by the theory of dynamical systems [199], that states specific rules to study nonlinear dynamics, i.e. through the analysis of the specific attractors, possibly chaotic, of the system under exam. This research in economics focuses on the idea that the observed randomness in financial time series, i.e. series of returns, price changes, spreads, etc., is not the expression of a stochastic dynamics but of a deterministic chaotic dynamics. We do not follow this line of re-

search, but it is important to remember that the concept of nonlinearity for financial variables has been also studied in the context of deterministic systems.

Another major feature of high frequency financial data is the irregular spacing in time. The question of how this salient property should be treated in an econometric model is not obvious. Indeed, the time between market events carries information and is a valuable economic variable serving as a measure of trading activity and affects price and traded volume behavior. Accounting for the timing of market events requires to consider the data statistically as *point processes*. Point processes characterize the random occurrence of single events along the time axis in dependence of observable characteristics and of the process history. This led to a big effort in modeling the dynamics of trade-to-trade durations. A popular model for durations is for example the ACD model, i.e. autoregressive conditional duration model, developed by Engle and Russel [94]. Other models used to reproduce irregularly spaced recordings are dynamic models for the intensity of events functions. Among such models we can find the autoregressive conditional intensity (ACI) model [220] and self-exciting intensity process like Hawkes process [136]. In this thesis we neglect the problem of having irregularly time recording for the variables under exam, for example prices, spreads, etc. Our main interest is instead to modeling the observed nonlinearities of price dynamics.

The purpose of the first part of the thesis is to investigate the importance of a nonlinear modeling for high frequency financial data describing large tick stocks. A brief overview of limit order books is given in chapter 2. In chapter 3 we show the empirical motivation that have led us to model price changes and bid-ask spreads by means of discrete processes. In chapter 4 we develop a family of Markov switching models and in chapter 5 we test the models against the empirical data by means of information-based metrics.

1.2 Market impact and the problem of optimal execution

The market impact is the average response of prices to trades, i.e. it is the effect of trading on the subsequent value of the price. When a trader buys a certain volume of a stock its price is pushed on average upward, instead when he sells the price is pushed on average downward. There are two interpretations about why the execution of orders impacts prices.

The first one describes the impact as the way by which the market processes the information in order to maintain its efficiency [42]. This is the standard view in economics, where perfect efficiency reflects perfect information processing. A standard definition of market efficiency, i.e. in the informational sense, is that the current price should be the best predictor of future prices, i.e. prices should be a martingale. Traders should process each new bit of information as it arrives moving the price to a new efficient equilibrium value. The main concept in economics is that the price moves because the agents in the market are asymmetrically informed about the value of the

traded security. The models for asymmetric information are described by a subset of agents that has superior private information [116, 153, 97]. Private information may consist of a signal about terminal security value, or more commonly, perfect knowledge of the security value. When such informed traders execute orders, a part of the private information is revealed to the market. If some traders really know the true price at some time in the future, i.e. the end of the day, the observation of an excess of buy trades allows the market, i.e. other agents like dealers or market makers, to guess that the price will move up and to change the price quotes accordingly. In this sense, information has progressively included in prices as a function of the observed order flow. Hasbrouck [134] has described this concept in this way: "orders do not impact prices, it is more accurate to say that the orders forecast prices ".

The second interpretation describes the impact as a mechanical phenomenon, i.e. the orders modify the available liquidity of the market causing a price shift. The market is a place where the agents randomly alter their demand/supply functions asynchronously. This framework is completely different from standard models in economics, because there are no assumptions about agents rationality. Such approach is mainly followed by the physics/mathematics community. From this prospective the flow of orders can be modeled as a random process, i.e. it is not determined by informed decisions of the agents in the market [82, 231, 71]. The models that made use of stochastic processes to describe the variables defining the limit order book dynamics, i.e. order arrivals, cancellations and the queues of volumes available for buying and selling, are called zero intelligence models [70]. According to such models the orders cause a price movement, as the result of their interaction with the limit order book.

We are not interested to understand if one approach is better than the other, but in this thesis we follow the mechanistic and phenomenological point of view. We focus our attention on two related mechanistic market impact models of the quantitative/physics literature: the model of Lillo and Farmer [96], and the model of Bouchaud et al. [43, 44]. Both models are developed in order to explain the apparent paradox deriving by an order flow strong correlated in time and the uncorrelated prices, i.e. the price efficiency. The long memory[2] of the sign of orders appears to be universal, it has been verified for stocks traded on the London Stock Exchange, the Paris Bourse and the New York Stock Exchange [164, 43]. It is also called "the long memory of supply and demand", because it implies that fluctuations in individual supply functions (leading to seller initiated transactions) and demand functions (leading to buyer initiated transactions) also exhibit long memory [166, 42]. If buying tends to push the price up and selling tends to push the price down, and we know that buying and selling are highly autocorrelated (and therefore predictable), how is it that prices remain uncorrelated and unpredictable? At this point, one can understand the importance of how the orders impact prices, i.e. the orders have to move prices preserving the efficiency of the price process. According to Lillo and Farmer [96], the market impact is permanent in time but depends on the state of the

[2] A long memory process is characterized by a slow decaying not integrable power law autocorrelation function.

market. The liquidity fluctuations, correlated with the order flow, act to suppress the otherwise permanent impact of orders and make the price diffusive. In contrast, Bouchaud et al. [43] describes the market impact as market state-independent and decaying as power law. The decaying factor is called the "bare "impact function or the propagator. The price at a given time is given by the sum of the impact of all the past trades weighted by the propagator function. In order to obtain a diffusive price, the propagator is a decaying power law function, defined by an exponent linked to the exponent describing the slow decay of the autocorrelation of the order signs. According to Bouchaud et al. [43], the propagator describes the mean-reversion of price due to the quote revisions of liquidity providers or market makers. The effect of trades on price shift is only transient and it disappears on long time scales. The seemingly trivial random walk behavior of price changes in fact results from a fined-tuned competition between two opposite effects, one leading to super-diffusion, i.e. the autocorrelation of the market order flow; the other leading to sub-diffusion, i.e. the decay of the bare impact function reflecting the long-range mean reverting nature of the limit order flow. On the other hand, as explained by Gerig [110] and Bouchaud et al. [42], these two models are equivalent in terms of price dynamics. Starting from the model of Lillo and Farmer [96] and assuming that the agents forecast the order flow using a long memory linear model, e.g. a FARIMA model, the price evolves like in the framework of Bouchaud et al. [43, 44]. In this thesis we consider the continuous time limit of the model of Bouchaud et al., developed by Gatheral [104], in order to study optimal strategies in the context of a nonlinear transient market impact.

A common problem for stock traders consists in unwinding large volumes of shares, which can comprise up to 20 % of the daily traded volume available on the market. These large institutional orders are often referred to as hidden orders or metaorders [42, 97, 238, 87]. In this thesis we choose to make use of the term metaorder. The optimal way of slicing the metaorder, given a price model and a cost functional, is called the optimal execution problem. As demonstrated in the context of a linear equilibrium model by Kyle thirty years ago [153], the optimal strategy for an investor with insider information on the fundamental price of an asset is to trade incrementally through time. This strategy allows the trader to minimize costs whilst also minimizing the revelation of information to the rest of the market. A correct model for the impact of the volume of trades on prices, i.e. the market impact, is a basic building block for a price model involved in designing optimal execution strategies. In the Kyle's model, for example, the price shift implied by trades is given by a linear function. The literature regarding optimal execution strategies focuses almost only on linear models of the market impact function. Problems of this type were investigated by Bertsimas and Lo [35], Almgren and Chriss [15, 16] and Obizhaeva and Wang [194].

However, a series of empirical studies [165, 43, 26] has clearly shown that the market impact, i.e. the expected price shift conditioned to the executed volume, is a strongly concave function of the traded volume, well approximated by a power law function. As observed and modeled by Bouchaud et al. [43, 44] and Obizhaeva and Wang [194], the price is also subject to a resilience effect. This resilience means that the mechanical price shift, due to the single trades, tends to disappear after a certain amount of time. These observations imply that the market impact is described more

properly by a function of the volume of trades and of the time elapsed since the execution. If we adopt the simplifying modeling assumption, proposed by Bouchaud et al. [209, 43], to factorize the impact function in a function of the volume and a function of time, we can describe these two terms like the instantaneous impact function and a transient, or decay, resilience function.

The model proposed by Bouchaud et al. [43], i.e. the propagator model, describes a price process where both the instantaneous market impact and the decay component, i.e. the propagator function, are nonlinear functions. In this thesis we are interested in the analysis of the optimal execution problem in the context of the propagator model. We consider the case of a concave instantaneous market impact function and a convex decay component. The main difficult of the analysis is due to the presence of the concave market impact function. The optimal strategy in the case of a linear instantaneous market impact and a convex decay component was described by Gatheral at al. [108].

The chapter 6 regards the general context of the optimal execution and price manipulations. The analysis of the propagator model is covered in chapters 7 and 8. The two chapters deal the same optimization problem by using different methods. In chapter 7 we analyze the problem by searching the optimal solution as the solution of a nonlinear Urysohn integral equation of the first kind. In chapter 8 we drop some assumptions on the properties of the optimal solution, used in chapter 7, and use a fully-numerical approach to search for the optimal strategy.

Part I

High frequency price dynamics

Chapter 2

Limit order books and tick size

In this chapter we introduce the institutional framework of trading on modern electronic financial markets. We introduce the basic concepts about the limit order book (LOB) and discuss the different types of order in section 2.1. We define the financial time series of interest for chapters 3, 4, 5, within the context of the limit order book in section 2.2.1. In sections 2.2.2 and 2.2.3 we illustrate the specific topics of *tick size* and *relative tick size*. Finally, in section 2.3, we give the details of NASDAQ stocks analyzed in chapters 3, 4, 5.

2.1 Limit order books and continuous auction

More than half of the markets in today's highly competitive and fast-paced financial world use a limit order book (LOB) mechanism to facilitate trade [130]. The Helsinki, Hong-Kong, Shenzhen, Swiss, Tokyo, Toronto, and Vancouver Stock Exchanges, together with Euronext and the Australian Securities Exchange, all now operate as pure LOBs; the New York Stock Exchange (NYSE), the National Association of Securities Dealers Automated Quotations (NASDAQ), and the London Stock Exchange (LSE) all operate a hybrid LOB. These markets differ in their details, but they all do their trading through a continuous double auction. *Auction* indicates that participants may place quotes (also called *orders*) stating the quantities and prices at which they are willing to trade; *continuous* indicates that they can update, cancel, or place new quotes at any time, and *double* indicates that the market is symmetric between buyers and sellers. Traders worldwide have real-time access to the LOB, providing buyers and sellers the ultimate microscopic level of description [45].

Some of the main fields of interest regarding the analysis of LOB dynamics include: optimal order execution strategies [194], design better electronic trading algorithm, and assessing market stability [10]. Investigations of LOBs have taken a variety of starting points, drawing on ideas from economics, physics, mathematics and statistics. There is no clear consensus on the best approach. For example there is a contrast between the approach normally taken in economics literature, in which models focus on the behavior of individual traders and present LOBs as sequential games [116, 153, 219], with the approach normally taken in the physics literature, in which order flows

are treated as random processes and techniques from statistical mechanics are used to explore the resulting dynamics [70].

The LOB is the list of all the buy and sell *limit orders*, with their corresponding price and volume, at a given instant of time. A limit order specifies the maximum (minimum) price at which an investor is willing to buy (sell) a certain number of shares (volume). At a given instant of time, all limit buy orders are below the best buy order called the *best bid price*, while all sell orders are above the best sell order called the *best ask price*. When a new order appears (say a buy order), it either adds to the book if it is below the ask price, or generates a trade at ask if it is above (or equal to) the ask price. We call all these *market orders*, even if technically they could also be *marketable* limit orders. It is important to recognize that this terminology, i.e. limit/market order, is used only to emphasize whether an incoming order triggers an immediate transaction or not. Limit orders, that fail to result in an immediate transaction, are stored as queues in the LOB. Such active orders can also be cancelled if the owner of the order no longer wishes to offer a trade at the stated price. The price dynamics is therefore the result of the interplay between the order book and the order flow.

The price gap between the best ask and the best bid is called the *bid-ask spread*. The liquidity [1] of the market is partially characterized by the bid-ask spread, which sets the cost of an instantaneous round trip of one share (a buy instantaneously followed by a sell, or the contrary). A liquid market is such that this cost is small [134]. Thus the spread plays a dominant role in the literature on market microstructure and stock trading [88]. Foucault et al. [102] argue that the bid-ask spread is the dominant parameter for the decision between different order type on stock markets. Traders can either be patient and submit limit orders or cross the spread and pay the bid-ask spread in order to obtain an immediate order execution. Empirical studies confirm that limit and market order submission strategies depend strongly on quoted spreads [131, 126, 201].

Figs. 2.1,2.2 show two snapshots of the first ten levels of the LOB relative to Apple Inc. (AAPL) and Microsoft Corporation (MSFT) traded at NASDAQ. The length of various rows represents the volume of shares of the queue of limit orders for a given price level. The two snapshots have a different structure, the first has price levels much more sparse than the second. Price gaps between quotes are nearly absent for the MSFT stock [90]. This leads to think that these stocks belong to different classes. This is indeed the case, the main features of these different kinds of stocks is the subject of the next section 2.2.

[1] Liquidity is difficult to define formally. Kyle [153] pointed out the three key properties of a liquid market to be tightness, i.e. the cost of turning around a position over a short period of time, depth, i.e. the size of an order flow innovation required to change prices a given amount, and resiliency, i.e. the speed with which prices recover from a random uninformative shock.

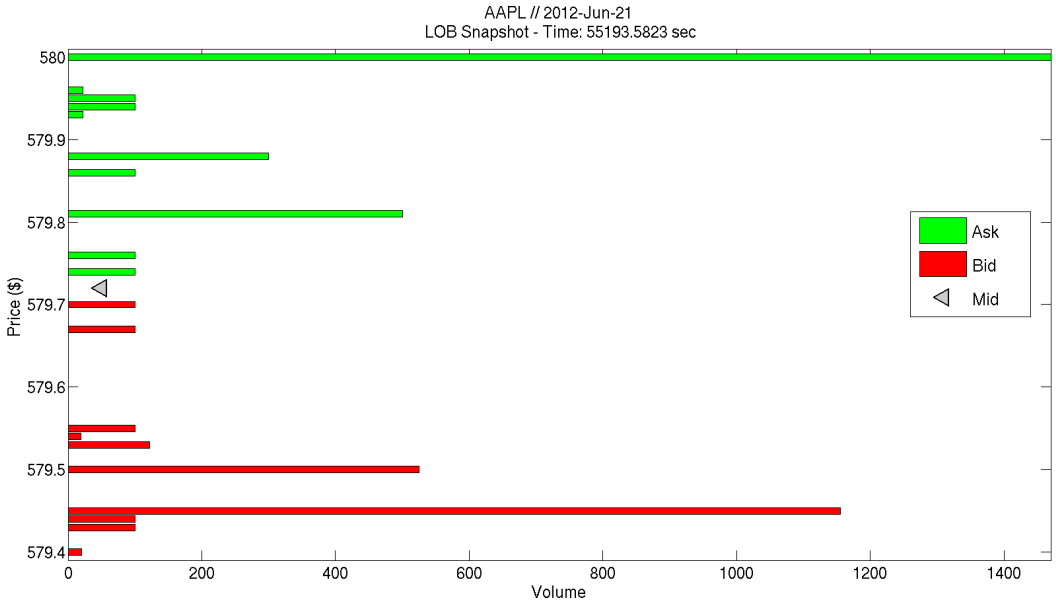

Figure 2.1: LOB snapshot relative to AAPL stock traded at NASDAQ market. The columns show the volumes of limit orders queues.

2.2 Tick size and time series of interest

In financial markets, the price of an order cannot assume arbitrary values but it can be placed on a grid of values fixed by the exchange. The tick size is the smallest interval between two prices, i.e. the grid step, and it is measured in the currency of the asset [23]. It is institutionally mandated and sets a limit on how finely prices may be specified. All price information is discretized by the tick size. Historically, the tick size of most securities has been consecutively reduced, resulting in tick sizes of 1/100th or smaller. This process is often referred to as decimalization [112, 137, 62, 170, 5]. The current tick size for stocks traded in US stock exchanges, such as the New York Stock Exchange (NYSE) or the National Association of Securities Dealers Automated Quotations (NASDAQ), is typically \$0.01. An argument for maintaining the tick size is that it serves to maintain a minimum level of profits for market makers and thus guarantees the provision of liquidity [176, 142, 41], but a too large tick size increases the transaction cost to investors by increasing the bid-ask spread. It is controversial whether a smaller tick size generally improves market quality.

2.2.1 Definition of time series

The main variables of interest of our work are: the mid-price, bid-ask spread, mid-price changes, mid-price log-returns and transacted volumes. These are the main variables that describe the complex dynamics of the limit order book. However, there are also other variables such the ask/bid side depth [70] and price gaps [95], that we do not use in our analysis of LOB dynamics. The behavior of recorded time series

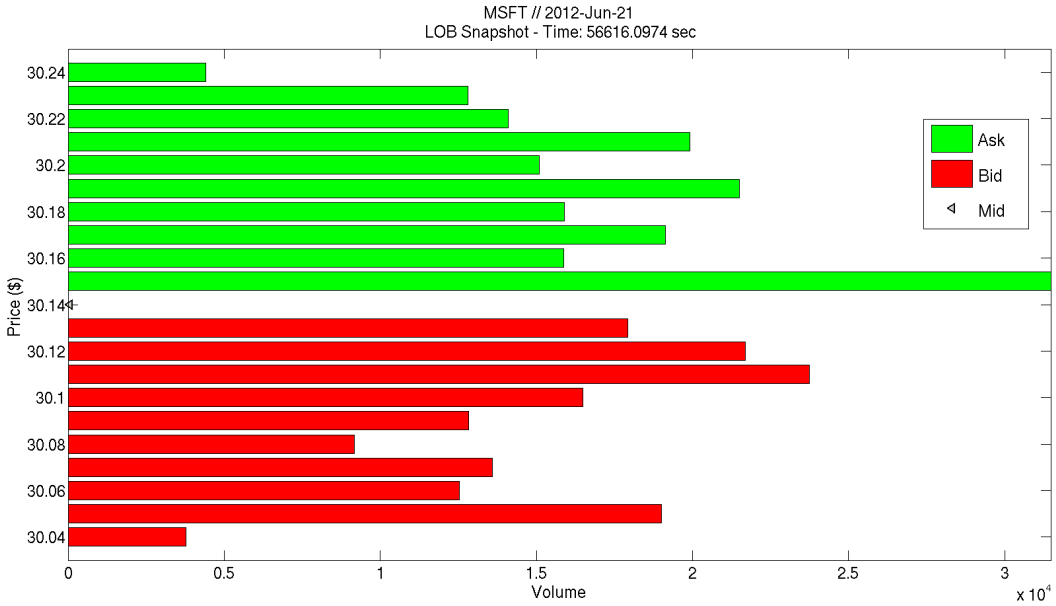

Figure 2.2: LOB snapshot relative to MSFT stock traded at NASDAQ market. The columns show the volumes of limit orders queues.

depends significantly on how it is sampled [182]. In this work we made use of two time scales: *continuous time scale* and *trade-by-trade time scale*. When the sampling is regularly spaced in time, i.e. with Δt seconds between successive samplings, such a time series is said to be sampled on a Δt-second continuous time scale. When the sampling times are chosen to correspond to trades, i.e the execution of a market order, the recording times may be spaced irregularly in time. The recorded times corresponding to trades can be subject to aggregation[2], i.e. when the interval of time between trades is lower than the resolution of the recording data system. Such a time series is said to be sampled on a trade-by-trade time scale. Hereafter the trade by trade time is considered to be an integer counter of events defined by the execution of a market order. The following definitions will be mainly used in chapters 3, 4, 5.

We define the following time series in trade by trade time:

- t_i is the time of i^{th} trade, $i \in \mathbb{N}$ is the transaction or trade time.
- $b(i) = b(t_i)$ and $a(i) = a(t_i)$ are respectively the best bid and ask prices just before the i^{th} trade.
- q_i is the volume, i.e. number of shares, of the transaction at time t_i.
- $p(i) = p(t_i) = (b(t_i) + a(t_i))/2$ is the midpoint price just before the i^{th} trade, $p \in \mathbb{N}$ is measured in units of half tick.

[2]It is important to point out that, since when a market order hits several limit orders, it results in several trades being reported in LOB's database, one possible choice is to aggregate together all such transactions and consider them as one trade if the time stamps of trades in the database are the same. We adopt this sampling technique.

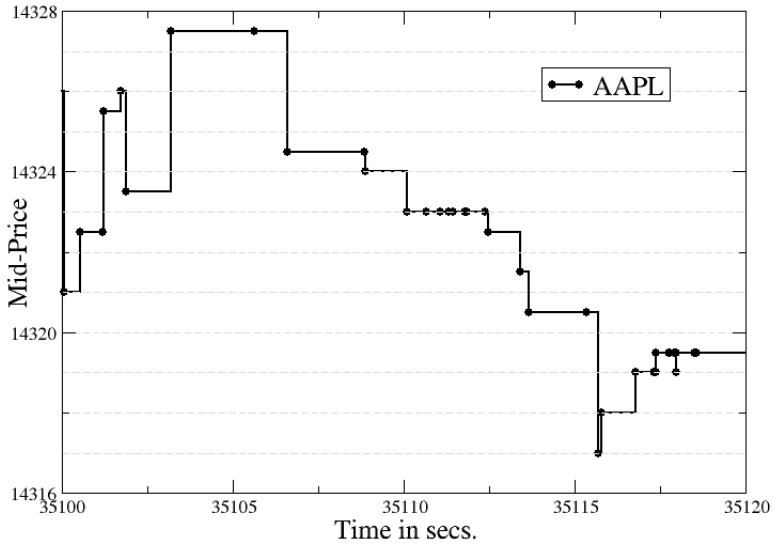

Figure 2.3: Midprice process for the stock AAPL is the black piecewise constant curve. The price is measured in units of one tick. The time is the number of seconds from the beginning of the trading day. Circles indicate executions of market orders. We can observe trades that do not cause price changes.

- $s(i) = s(t_i) = a(t_i) - b(t_i)$ is the spread just before the i^{th} trade, $s \in \mathbb{N}$ is measured in units of tick.
- $\Delta p(i) = \Delta p(t_i) = p(t_{i+1}) - p(t_i)$ is the price change caused by the i^{th} trade, $\Delta p \in \mathbb{Z}$ is measured in units of half tick.
- $r(i) = r(t_i) = \log(p(t_{i+1})) - \log(p(t_i))$ is the log-return, $r \in \mathbb{R}$.
- $\Delta p(i,n) = p(t_{i+n}) - p(t_i)$ is the price change caused by n consecutive trades, n is the trade time scale at which we observe the price change process.
- $r(i,n) = \log(p(t_{i+n})) - \log(p(t_i))$ is the log-return caused by n consecutive trades.

We want to study the price process also in continuous time. To this end we define the midprice process $p_c(t)$ assuming that the price between two transactions is given by the midprice just before the second transaction. This defines a piecewise constant function like the one shown in Figure 2.3.

- $p_c(t) = p(t_{i^*})$, where $t_{i-1^*} \leq t < t_{i^*}$ is the time between the two subsequent transactions $i - 1^*$ and i^*.
- $\Delta p_c(t, \Delta t) = p_c(t_{i^*+n^*}) - p_c(t_{i^*})$, is the price change observed by sampling the time series $p_c(t)$ at a time scale Δt. This change is caused by n^* consecutive

trades. The number of trades n^* is a stochastic variable for each fixed value of the time scale Δt.

- $r_c(t, \Delta t) = \log(p_c(t_{i^*+n^*})) - \log(p_c(t_{i^*}))$, is the log-return caused by n^* consecutive trades.

2.2.2 Literature review on tick size

A large part of the studies about tick size present in the literature are case studies of the impact of a reduction of tick size on market quality, i.e. on microstructural quantities like the narrowing of the bid-ask spread [170] or liquidity provision [117, 12]. A part of the literature is more related on how the investors actually use the price resolution allowed by the tick size. This can led to an effect of clustering of prices, i.e. an unevenly use of the available price grid. Other studies focused on statistical properties of price fluctuations [197, 191, 155] and on the connection between bid-ask spread and midprice dynamics [84, 250, 215].

Onnela et al. [197] study the effect of changes in tick size, enabled by the decimalization process, on asset log-returns. They analyze a set of NYSE and TSE (Toronto Stock Exchange) cross-listed stocks that were traded under different tick sizes. The data were daily closing prices from Jan-1-1990 to Jun-30-2003. They show that investors do not use all price fractions uniformly as allowed for by the tick size, leading to a clustering of prices on certain fractions, a phenomenon that could potentially affect the way returns are distributed. This phenomenon persists after decimalization. They observed that approximately 57% of cases exhibit price clustering such that the effective tick size deviates from the nominal tick size. In this study the tick-to-price ratio, appears to be indicative of the zero returns frequency. They conjectured that large effective ticks lead to a distortion of the shape of return distribution, and this effect should be particularly strong when the price of stock is low, i.e. when tick-to-price ratio is high.

Munnix et al. [191] demonstrate that the tick size has a large impact on the structure of financial return distributions. They analyze a basket of stocks from the S&P 500 index ranging from 1 min to 1 day frequency during the first half of 2007. They find returns clustering at 1 min frequency, and observe that the discrete distribution of price changes could lead to think that the transition from integer price changes to relative price changes, i.e. returns, remove the discretization from the distribution. A closer analysis instead reveals that the discretization effect are still visible when considering returns. They argue that the discretization affects returns on any time scale. They perform an approximate analysis that reveals a sort of mapping between the discrete distribution of price changes and the distribution of returns. They decompose the set of returns according to the absolute price changes, i.e. one value of price change corresponds to a specific set of returns. Their computations lead to the conclusion that the width of this sets are proportional to the absolute value of price changes, while the distance between their centers remains almost constant. In this way the sets of values of returns are increasingly overlapping for larger values of absolute price changes. From their viewpoint the discretization is only visible

for small absolute price changes, i.e. one could see an unusual distortion of return distribution near its center. Moreover they find that the shape of the distribution of normalized returns compared to the underlying normalized price changes are quite similar for time scales ranging from 5 min to 1 day. According to Munnix et al. [191] the meaning of clustering is that the distribution of returns is defined on specific sets of the real line, i.e. we do not have a smooth distribution like the Gaussian or Lévy distributions. This effect is less and less visible if we have a large number of possible different values for price changes, because we have the overlap of the different sets.

Two alternative hypothesis were given by the economic community for clustering: natural clustering [129, 198] or collusion [61]. Harris [129] studied the frequency distribution of the integer portion of CSRP daily closing price stocks for the years 1963 to 1987, including NYSE, AMEX and NASDAQ stocks. In this case the minimum ticks size ranged from \$1/8 to \$1/16, and the tick size was smaller for stocks with lower prices. He argued that stock price clustering is pervasive and that clustering distributions from the mid-nineteenth century appear very similar to those observed in the late twentieth century. Clustering increases with price level and volatility and occurs if traders use discrete price sets to simplify their negotiations. He claimed also that clustering must affect price changes distributions and bid/ask quote distributions. Collusion instead refers to the idea that market makers quote prices only in certain fractions in order to increase bid-ask spreads. Christie and Schultz [61] show that many NASDAQ stocks exhibit a paucity of odd-eights quotes and quote prices mainly in even-eights. Bessembinder [36, 37, 39, 38] provides empirical evidence on relations between trade execution costs and price rounding practices on the NYSE and NASDAQ. His results indicate that higher execution costs are associated with the rounding of quotations and trade prices, and finds that the effect of clustering on trading costs decreases as the tick size decreases.

The empirical analysis of price changes clustering for large tick stocks is the main topic of the chapter 3. The specific form of the distribution of price changes will be the motivation of the Markov modeling approach shown in chapter 4.

2.2.3 Relative tick size

The studies reported in section 2.2.2 regard more the behavior of stocks traded in markets defined by a fixed value of the tick size. More recently, the concept of relative tick size has been defined in the market microstructure literature [195, 252]. Due to its recent introduction, the same concept can be referred in different ways. The term *relative* can be also referred to as *perceived* [83, 84] or *effective* [197]. However, each of them describes the same behavior, i.e. a different dynamics of stocks in terms of price and spread dynamics for a given tick size [90]. We make use of two definitions of the relative tick size:

- $T_r = 1/\mathbb{E}[p_t]$ is the tick-to-price ratio, where the trade price p_t is measured in units of 0.01\$. This definition follows O'Hara et al. [195]

- $T_s = \#\left[s\left(i\right) = 1\right]/N_t$ is the fraction of times the spread is equal to 1 tick, N_t is the total number of trades in the data sample. This definition follows Dayri et al. [83].

These empirical properties are useful to obtain a ranking between stocks belonging to the same market. We will make use of such definitions in what follows.

Why the relative tick size is important? The literature regarding the relative tick size is relatively poor and regarding mainly U.S. markets. There are at least two motivations for this kind of classification, i.e. a rough division between large and small tick stocks, the first one is institutional and the second derives from the differences found in the high frequency price dynamics. Recently the SEC (Securities and Exchange Commission) have posed the issue of the possible impact that decimalization, i.e. a tick size equal to 1 cent, has had on liquidity for small and middle capitalization company securities [5]. The commission requested comments on whether public price discovery and execution quality have suffered, and specifically questioned whether the minimum pricing increment for lower priced stocks should be reduced or increased. This request was motivated by the precipitous fall in IPOs (Initial Public Offering) and the decreased volumes of trading in small capitalization stocks. Proposals to answer the above question usually involve implementing pilot studies of different tick sizes regimes, this is a complex and costly approach. O'Hara et al. [195] believed that there is a more direct way to address this issue, they designed an empirical study based on the concept of the relative tick size. They analyzed a basket of stocks traded on NYSE between May and June 2012. This research design exploits the fact that the relative tick size, in this case defined as the tick size relative to the stock price, is not uniform across stocks, but can vary substantially depending upon stock price levels. By matching stocks with large relative tick sizes to a control sample of similar stocks with smaller relative tick size, it is possible to isolate the specific effects of tick size on liquidity and trading environment. The main findings are that high-frequency trading firms that operate as market makers on the NYSE take on a more prominent role in liquidity provision for stocks with larger relative tick sizes: spending more time at the quote, improving market-wide prices, and increasing their participation in trading. A larger relative tick size does not, however, seem to attract more overall trading volume from investors to the stocks.

The literature regarding the impact of the relative tick size on price dynamics is almost absent. Robert and Rosenbaum [215] proposed a model that accommodates the assumption of a continuous efficient price with the inherent properties of ultra high frequency transaction data, i.e. price discreteness and irregular temporal spacing. A specific parameter quantifies the aversion to price changes of the market participants. When the aversion to price change is high we are in presence of a large tick stock. Their approach consists in designing a stochastic mechanism for deriving the transaction prices from the latent efficient price. The main result of the model is that it is able to allow for discrete price process, a bid-ask bounce and an inverse relation between duration and volatility. Moreover, Dayri et al. [83] observed that the parameter, used by Robert and Rosenbaum, quantifying aversion to price changes leads to the same ranking of assets obtained by the T_s spread-based criterion.

Table 2.1: Sample statistic for log-returns

Stock	time scale	mean	std.dev.	skewness	ex.kurtosis
AAPL	1 trade	6.602e-08	7.553e-05	0.01434	5.268
	1 sec	6.641e-08	8.043e-05	-0.00105	22.16
AMZN	1 trade	6.497e-08	1.484e-04	0.0442	8.177
	1 sec	4.590e-08	1.097e-04	0.3164	41.25
MSFT	1 trade	1.038e-07	1.190e-04	0.0112	7.773
	1 sec	5.949e-08	8.406e-05	-0.0146	50.50
CSCO	1 trade	1.173e-07	1.427e-04	0.00505	7.008
	1 sec	5.587e-08	9.791e-05	0.1285	50.34

We choose to follow a completely different approach, starting from studying the impact that the relative tick size, as defined above, can have on stock returns and price changes distribution. This analysis is performed in chapter 3. From this empirical analysis we develop a model specifically designed for large relative tick size stocks in chapters 4, 5. We do not use a model defined by an efficient price plus a microstructural noise. We study a model that describes directly the observed discrete price process. In the following, we drop the adjective relative and refer to this kind of stocks as large tick stocks.

2.3 NASDAQ data

In chapters 3 and 4 we study high frequency data of highly liquid stocks traded at NASDAQ market in the period from 01/07/2009 to 31/08/2009 (42 trading days). We analyze the stocks: Apple Inc. (AAPL), Amazon (AMZN), Microsoft Corporation (MSFT), Cisco Systems (CSCO). Our data contain time stamps corresponding to order executions, trade prices, bid-ask quotes, size of trading volume and direction of trading. The time resolution is millisecond. The trading activity at NASDAQ starts at $9:30$ and ends at $16:00$. We decide to discard all transaction data corresponding to first and last 6 minutes of the day. During these minutes we observe bursts of trading activity and an abnormal high price fluctuations that could affect the statistical analysis of returns and price changes distributions.

We report in table 2.1 some sample statistic about log-returns corresponding to the smallest time scales studied in this work. When we observe prices in continuous time, the empirical returns distribution is more fat-tailed than that defined in trade time. The increase of kurtosis can be explained if we think to price process as a subordinated random process. This argument will be reviewed in chapter 3.

We apply to the stock's data the definitions of relative tick size given in section 2.2.3. We can observe in table 2.2 that these two measures divide the stocks in the same manner in two groups: AAPL and AMZN are small tick size stocks, instead MSFT and CSCO are large tick size stocks. It is important to observe that the two

Table 2.2: Effective tick size for NASDAQ stocks. [a] 42 days of transactions. [b] mean time value in sec between 2 trades. [c] measured in basis points.

Stock	tick size	# Trades N_t^a	Duration [b]	T_r [c]	T_s	Class
AAPL	0.01$	918294	1.037	0.64	0.256	SMALL
AMZN	0.01$	530076	1.797	1.2	0.243	SMALL
MSFT	0.01$	532795	1.788	4.1	0.932	LARGE
CSCO	0.01$	420963	2.263	4.9	0.932	LARGE

measures lead to the same classification between large and small tick assets, as also observed by Dayri et al. [83] on a different basket of stocks.

As we will see in the following, these two classes are different from the point of view of the statistical properties of their price changes and returns distributions from small to large time scales of observation. Starting from these empirical facts we develop models for large tick stocks. Then we propose a multiscale model selection method based on the Jensen-Shannon distance [124] in order to select the model that is able to better reproduce the distribution of price changes at different time scales. We study the price process at different time scales and compute the Jensen-Shannon distance between the original dataset and different models, showing that the coupling between spread and returns is important to model return distribution at different time scales of observation, ranging from the scale of single transactions to the daily time scale.

Chapter 3

Dependence of stock returns and spread distribution on the tick size

3.1 Introduction

Tick size can affect prices in a direct way on different time scales, starting from the microstructural scale to the daily scale. In this section we analyze the midprice process, i.e. the dynamics of midpoint between bid and ask quotes, in transaction time and in continuous time. We want to study the scaling of the distributional properties of price fluctuations at different time scales, starting from the smallest time scale, e.g. price changes and log-returns caused by 1 transaction. In this way we can see the connection between high frequency dynamics of prices, i.e. 1 sec or 1 min dynamics, and low frequency dynamics, i.e. 1 hour dynamics. The basic observation is that at the smallest time scale the distributions of returns are very far from Gaussian or Levy stable distributions, that are instead used to model price fluctuations at higher time scales [46, 135, 109]. The return distribution at the smallest time scales strongly depends on the value of the tick-to-price ratio. As it is known in the literature, the value of the tick size is not the best indicator for understanding and describing the high frequency dynamics of prices. The tick-to-price ratio, defined as T_r in section 2.2.3, is one of the definitions of the notion of a relative tick size, introduced in order to account and quantify the different behavior of price fluctuations. Another useful definition, referred to as T_s in section 2.2.3, is based on the bid-ask spread. In this case the measure is given by the frequency the spread is equal to one tick and we have a large tick size if the spread is almost always equal to one tick. Usually these measures of the relative tick produce the same ranking between different securities.

The key observation is that for large tick stocks the price changes are anomalous clustered on the grid of the possible integer values that they could assume. Specifically, we find that even price changes are more populated than odd values. This property is found to hold from small to high time scales. Instead for small tick size stocks the clustering of price changes is not present. The high frequency dynamics of price for a large tick stock is characterized by the presence of clustering. A similar property has been reported in literature [129, 197] for daily closing price series. The

presence of clustering affects also the distribution of returns for large relative tick size stocks, while this effect is negligible for small tick stock.

We want to quantify empirically the distortion of the shape of distributions of price changes and returns as a function of the relative tick size, measured by the tick-to-price ratio or by the frequency of bid-ask spread equal to 1 tick. We expect that, after a certain time scale of aggregation, the shape of distributions become independent from the relative tick size of the stock. The distortion can be characterized by measuring how far the distributions are from the Gaussian.

Moreover, we analyze the differences regarding distributions of the bid-ask spread between large and small tick stocks in transaction time. Such differences will be useful in order to develop a stochastic model for price dynamics of large tick stocks in the next chapter.

3.2 Empirical analysis

In this section we study the role of the relative tick size on the distributional properties of price changes and log-returns at different time scales. We start with the study of the shape of distributions of price changes $\Delta p\left(i,n\right)$ and log-returns $r\left(i,n\right)$ as a function of the value of the tick-to-price ratio. As defined in section 2.2.1, the trade time is described by the integer i and the time scale of observation is described by the integer n. In this qualitative discussion we refer to price changes and returns computed in trade time and continuous time. We want to show that the effect of a discrete tick size is more substantial for a high tick-to-price ratio.

The first important observation is the presence of price changes clustering when we observe the price change process in trade time or in continuous time. The price change clustering is the phenomenon for which we have an uneven use of price fractions of the price grid. Clustering is not observable at the scale of 1 transaction, but becomes significant after few transactions for large tick stocks. In fig. 3.1 we show the histogram of price changes caused by 1 transaction for large and small tick stocks. The main differences are that for large tick stocks the price changes are almost always zero and the probability of having price changes of absolute value grater than 1 tick is almost zero. The difference between the two classes of stocks becomes clear when we observe price changes at a greater transaction time scale. In fig. 3.2 we show the histogram of price changes at an aggregation scale $n = 128$ for large and small tick stocks. A large tick stock has a distribution of price changes in which odd values are significantly less populated than even values. Our empirical observations indicate that the process $\Delta p\left(i,n\right)$ shows clustering for each value of the transaction time scale greater than one, i.e. $n > 1$, in the case of large tick stocks. For example if we observe the process $\Delta p\left(i,n = 8192\right)$ the clustering is still present and 8192 transactions are a significant part of the total transactions that we could have in one day of trade, e.g. in the case of MSFT they correspond to an average execution time of four hours. Therefore, this effect is visible not only at high frequency time scales. In the next chapter we propose a price model that can account for price change clustering, starting from the price dynamics at the scale of single transactions.

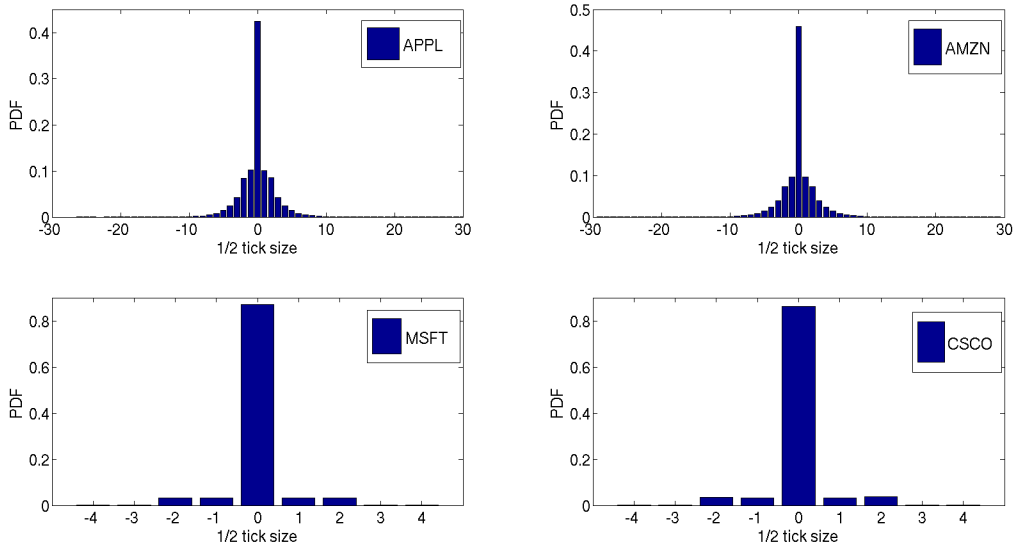

Figure 3.1: Histograms of price changes measured in trade time caused by one transaction, i.e. $n = 1$, for AAPL, AMZN, MSFT and CSCO stocks.

The effect of price changes clustering is not present at all in the case of a small tick-to-price ratio. From the smallest time scale to the largest, i.e. from 1 to 8192 transactions, we observe a usual occupation of even and odd levels of price changes. When we refer to a usual occupation we mean an absence of systematic differences in populations of price changes and a presence of a smooth discrete distribution like a binomial or a Poisson distribution. The differences in the shape of price changes distributions between small and large tick size stocks are clear in fig. 3.2. The observation of clustering at a daily time scale is already known in literature [197] but it is not clearly connected to a measure of the relative tick size. Our observations, instead, connect this property of prices directly to the relative tick size. Such observations do not change if we observe the price at a continuous time scale, as is shown in figs. 3.3, 3.4. We conclude that we do not have a universal shape of distributions of price changes, but we have a dependence on a relative tick size, measured by T_r or T_s defined in section 2.2.3.

There is a growing consensus that distributional properties of returns are quite universal, i.e. the shape of the distribution is the same for all the stocks, especially for relatively large time scales [69]. How can we reconcile the empirical observation of price change clustering with a universal shape of returns distributions ? The presence of a discrete tick size has an effect on the distribution of returns. Our empirical analysis shows that return distributions are affected by an effect of discretization and by returns clustering[4]. For small tick size stock only the discretization effect is present,

[4]Notice that for returns the discretization effect is different from clustering: discretization is a consequence of the fact that price is defined on a grid, while clustering denotes the preference for some price variations over others.

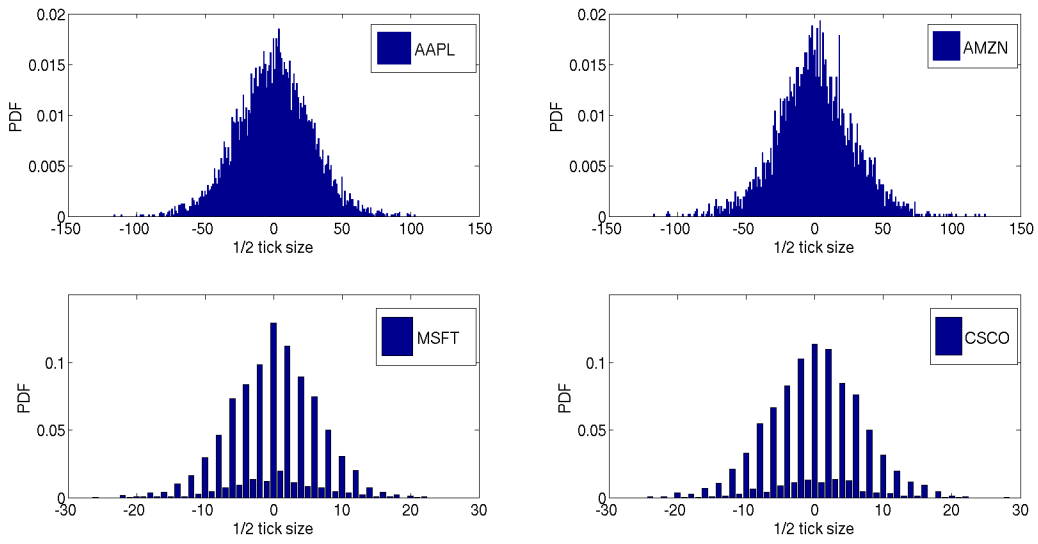

Figure 3.2: Histograms of price changes measured in trade time at an aggregation scale $n = 128$, for AAPL, AMZN, MSFT and CSCO stocks. The effect of clustering for MSFT and CSCO stocks is clearly visible.

while for large tick stock there is the additional effect of returns clustering. The effect of discretization is less present as the time scale increases, instead returns clustering is less present as the tick-to-price ratio decreases. We want to stress the idea that the presence of discretization, coupled with returns clustering, for an high tick-to-price ratio disappears at time scales n higher than that relative to a low tick-to-price ratio.

For small tick size stocks discretization effects on returns are visible at time scale of one transaction, see fig. 3.5, but when $n \approx 128$ this effect disappears. Instead for a large tick stock at the time scale of $n \approx 128$ the effect of discretization and clustering are present as we can see in fig. 3.6.

We want to stress the idea that the presence of discretization for an high tick-to-price ratio disappears at time scales n higher than that relative to a low tick-to-price ratio. Such observations are the same for our empirical observations in continuous time, as we can observe in figs. 3.7, 3.8.

Our working hypothesis is that we should find some time scale in trade or continuous time at which the discretization of returns disappears and we could find a universal shape for distributions of price returns. This means that we should study the properties of scaling of the distributions of returns. We performed this analysis by means of the empirical hypercumulants Λ_q of distributions and the tail exponent α describing the asymptotic power-law behavior of distributions. Our empirical observations imply that such distributions are not universal, i.e. as stated above clustering affects also large time scale of observations in the case of large tick stocks.

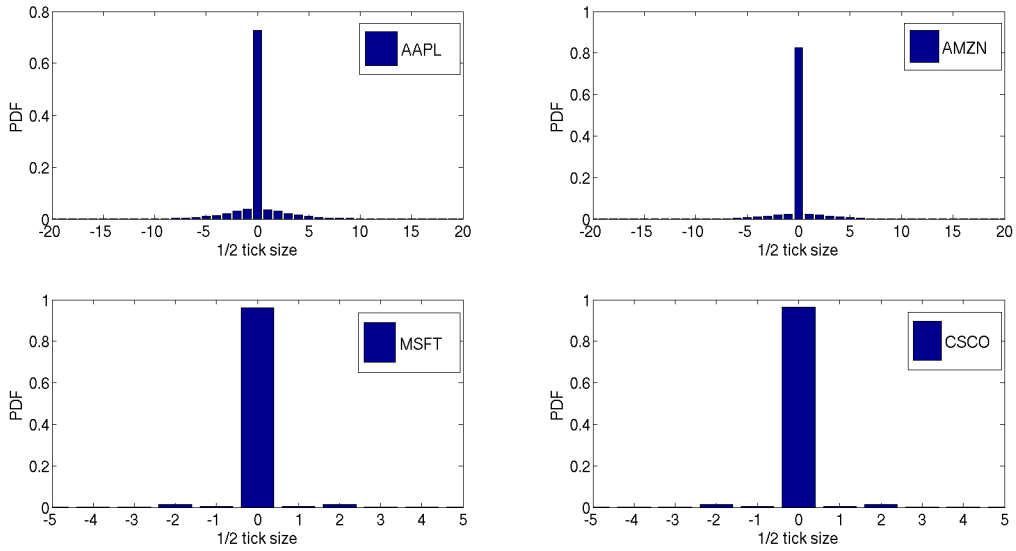

Figure 3.3: Histograms of price changes measured in continuous time for a time scale $\Delta t = 1$ sec, for AAPL, AMZN, MSFT and CSCO stocks.

3.2.1 Hypercumulants and tail exponents

The distribution of returns is one of the most basic characteristics of financial markets and many papers have been devoted to it. The Gaussian assumption, proposed the first time by Bachelier [25], for the distribution of returns is widely used in theoretical finance. In his pioneering analysis of cotton prices, Mandelbrot [180] observed that the process of returns is not a Gaussian process. Motivated by the pronounced tails of the return distribution and the stable functional form, i.e. from 1 day up to one month of observation, Mandelbrot proposed that the distribution of returns is consistent with a Lévy stable distribution. More recently, the availability of high frequency data has facilitated the difficult task of studying the tails and asymptotic behavior of return distributions. The main purpose of more recent research was to investigate the distribution of returns from time scales of the order of 1 minute, i.e. high frequency time scale, to the monthly time scale. The interest in the analysis of the shape and asymptotic behavior of return distributions came from the central limit theorem. The central limit theorem allows to connect the distribution of the high frequency returns to the distribution of returns to the daily or monthly time scales, in the hypothesis of i.i.d returns. This last assumption is certainly not true, due to the clustering of volatility present at almost all time scales of observation. This last observation implies that the analysis of the return distributions at different time scales is still an open field of research. Recently, several authors have established the the tails of stock returns and indices are well described by a power-law with exponent α in the range $3 - 5$ for time scales ranging from 1 minute to some days [46, 118, 205]. For longer time scales the results are consistent with an asymptotic power-law distribution and a slow convergence to a Gaussian behavior. We follow this line of research considering

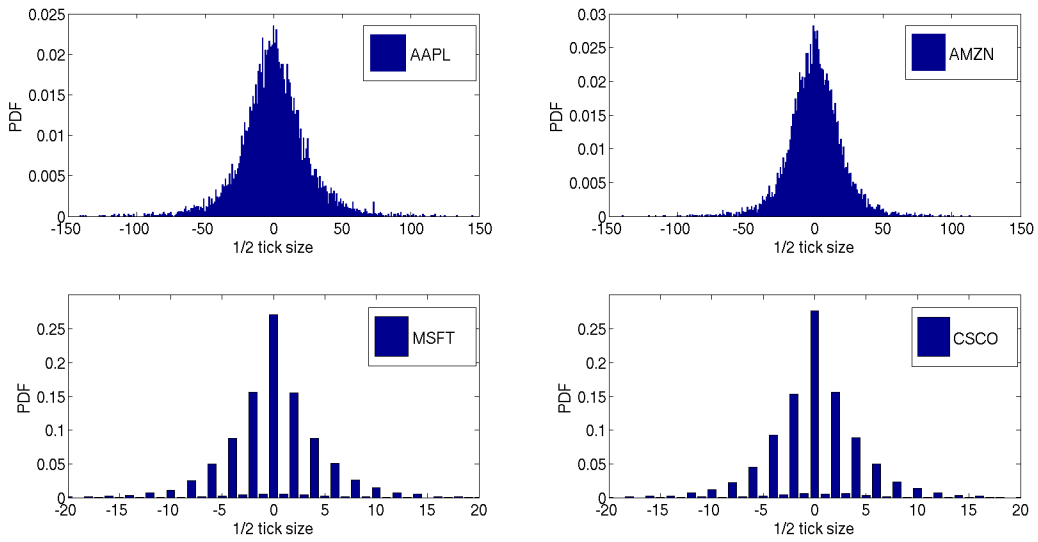

Figure 3.4: Histograms of price changes measured in continuous time for a time scale $\Delta t = 120$ sec, for AAPL, AMZN, MSFT and CSCO stocks. The effect of clustering for MSFT and CSCO stocks is clearly visible.

as a parameter, that can affect the distributions, the relative tick size. This analysis is performed by using two quantities: the hypercumulants of the distribution of returns, measuring the shape of the distribution, and the exponent of the power law tail, which determines the properties of returns under aggregation. We measure this two quantities for returns computed in continuous time and trade time scales. Our purpose is to study the scaling of such distributional properties at intra-daily time scales for small and large tick stocks. The analysis in continuous time ranges from the time scale of 1 sec to the time scale of two hours. The analysis in trade time ranges from the time scale of 1 trade to 8192 trades.

Methods:

In order to compare the behavior of distributions for different time scales, i.e. n for trade time or Δt in continuous time, we define a normalized return g:

$$g(i,n) = \frac{r(i,n) - \langle r(i,n) \rangle}{\sqrt{\langle r^2(i,n) \rangle - \langle r(i,n) \rangle^2}}. \tag{3.1}$$

The definition for the continuous time case is similar. We analyze the scaling by the moments defined by a fractional index q, i.e. the hypercumulants [46, 118], of the distributions of normalized returns $g(i,n)$:

$$\Lambda_q(n) = \mathbb{E}\left[|g(i,n)|^q\right], \tag{3.2}$$

in this way this quantity is defined as a function of the time scale n.

We estimate also the tail exponent α for the normalized returns g. This exponent describes the asymptotic power-law behavior of probability density functions in the

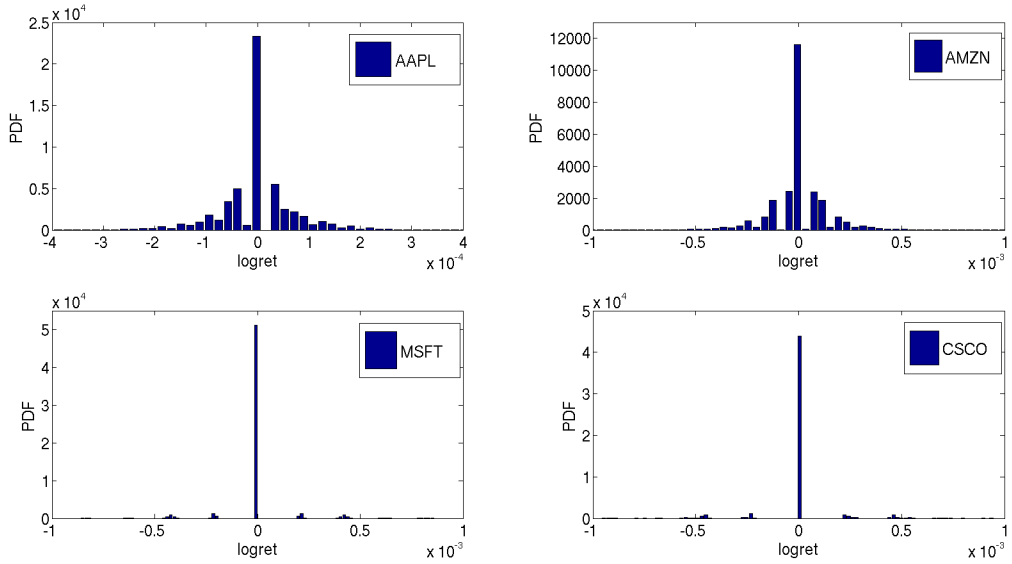

Figure 3.5: Histograms of log-returns measured in trade time caused by one transaction, i.e. $n = 1$, for AAPL, AMZN, MSFT and CSCO stocks.

following way:

$$P(x) \sim x^{-(1+\alpha)}, \tag{3.3}$$

where $\alpha > 0$. This estimate gives information on the existence of the moments of a given distribution. A necessary condition for the q^{th} moments to exist is that the distribution density $P(x)$ should decay faster than $1/|x|^{q+1}$ for $|x|$ going towards infinity, then all the moments such that $q > \alpha$ are infinite. The asymptotic behavior of the density $P(x)$ is also connected to properties of random variables under summation. Consider the sum $S_m = \sum_{i=1}^{m} x_i$ of independent identically distributed (i.i.d) random variables x_i. If the x_i's have finite second moments, the central limit theorems holds and S_m is distributed as a Gaussian in the limit $m \to \infty$. For example, this is the case when the asymptotic power-law behavior is described by $\alpha > 2$. If the random variables x_i are characterized by a distribution having asymptotic power-law behavior like that in eq. 3.3 where $0 < \alpha < 2$, then S_m converges to a Lévy stable stochastic process of exponent $0 < \alpha < 2$ in the limit $m \to \infty$. It is interesting to study the exponent α of the convolution of a finite number of variables. The convolution of a finite number of variables, described by a power-law asymptotic behavior with $\alpha > 0$, is still described by an asymptotic power-law defined by the same α. In presence of an infinite sample of i.i.d. asymptotic power-law distributed data, the exponent α, measured at different aggregation time scales, should be a roughly constant function of the time scale. Also in presence of a finite sample, if we had at the smallest time scale an asymptotic behavior described by $\alpha < 2$ with i.i.d. returns, we should observe a roughly constant scaling in function of the aggregation time, converging to a Levy distribution of exponent α. If, instead, we had at the smallest time scale an asymptotic behavior described by $\alpha > 2$ with i.i.d. returns, we should observe at the

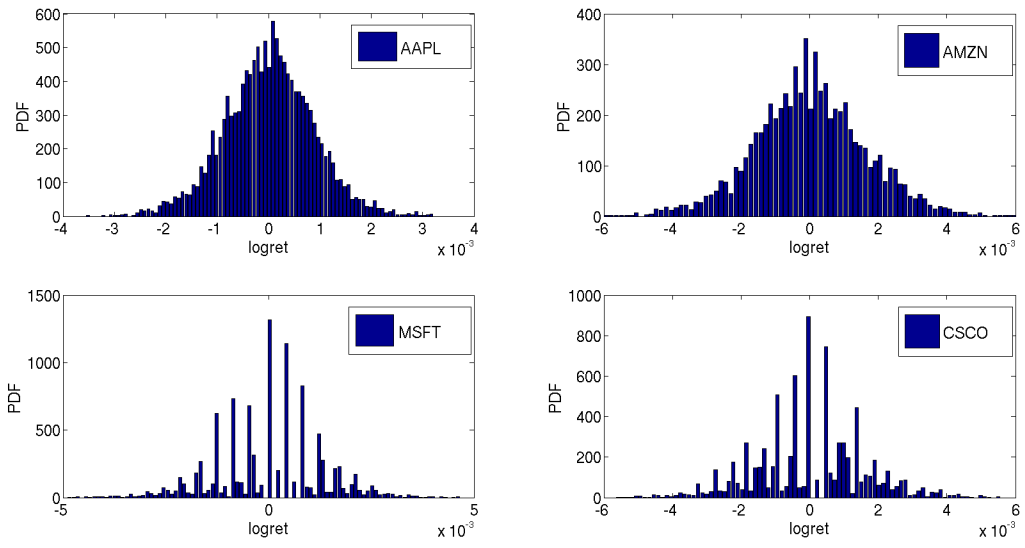

Figure 3.6: Histograms of log-returns measured in trade time at an aggregation scale $n = 128$, for AAPL, AMZN, MSFT and CSCO stocks. The effect of clustering for MSFT and CSCO stocks is clearly visible.

beginning a roughly constant function and than a slowly convergence to the Gaussian, with the consequent disappearance of the power law behavior. This effect for $\alpha > 2$ is due to the finite length of the data sample. The slow convergence to the Gaussian makes difficult to sample the power law tails. Therefore, from the scaling of α with the time scale of aggregation is possible to observe qualitatively the presence or less of time dependencies between returns from the smallest time scales [95] to time scales where is still possible to sample the power-law tails. This is usually possible for a time scale of aggregation that is small respect to the total length of the sample.

A common problem when studying a distribution that decays asymptotically as a power-law is how to obtain an accurate estimate of the exponent characterizing its asymptotic behavior. We use a method developed by Clauset et al. [64] in order to estimate the exponent α and the value of x ,i.e. x_{min}, beyond which we have the power-law behavior. Their approach combines maximum-likelihood fitting methods with goodness-of-fit tests based on the Kolmogorov-Smirnov statistics and likelihood ratios. They fit a Pareto tail distribution on the tails of the empirical distributions under exam, normalizing the data for which $x > x_{min}$

$$P(x) = \frac{\alpha}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-(1+\alpha)}, \tag{3.4}$$

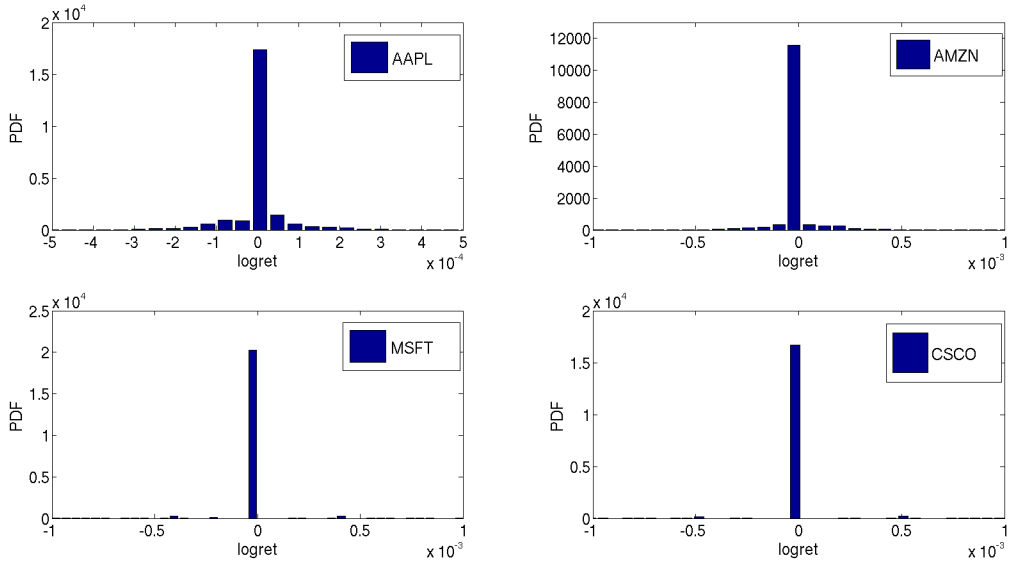

Figure 3.7: Histograms of price changes measured in continuous time for a time scale $\Delta t = 1$ sec, for AAPL, AMZN, MSFT and CSCO stocks.

where x_{min} is for us the lower bound of power-law behavior and $\alpha > 0$. They use the well-known Hill [140] maximum likelihood estimator:

$$\hat{\alpha} = m \left[\sum_{i=1}^{m} \ln \frac{x_i}{x_{min}} \right]^{-1}, \tag{3.5}$$

where $x_i,\ i = 1, \cdots, m$, are the observed values of x such that $x_i > x_{min}$. The Hill estimator is known to be asymptotically normal and consistent, i.e. $\hat{\alpha} \to \alpha$ in the limit of large m. In order to have a good estimate of α we need a reliable estimate of the lower bound x_{min}. The most common ways of choosing $\hat{x}_{min}$ are either to estimate visually the point beyond which the distribution function becomes roughly linear on a log-log plot, or to plot $\hat{\alpha}$ as a function of $\hat{x}_{min}$ and identify a point beyond which the value appears relatively stable. These approaches are clearly subjective and can be sensitive to noise or fluctuations in the tail of the distribution. Clauset et al. [65] proposed a new approach for estimating x_{min}. The value of $\hat{x}_{min}$ is determined choosing the value that makes the probability distribution of the measured data and the best fit power-law model as similar as possible above $\hat{x}_{min}$. There are a variety of measures for quantifying the distance between two probability distributions, but for non-normal data the commonest is the Kolmogorov-Smirnov or KS statistic [212], which is simply the maximum distance between the CDFs of the data and the fitted model

$$D = max_{x \geq x_{min}} |S(x) - P(x)|. \tag{3.6}$$

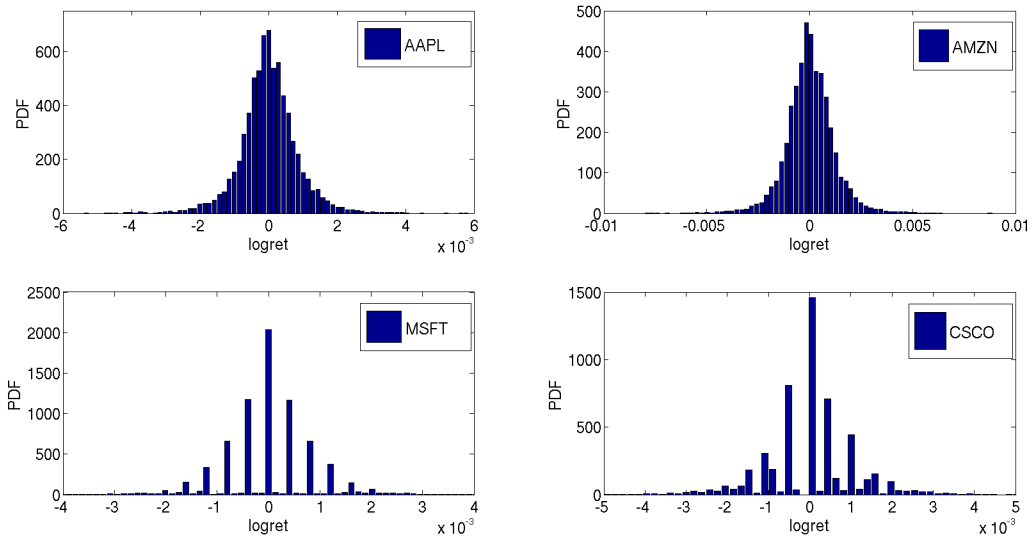

Figure 3.8: Histograms of price changes measured in continuous time for a time scale $\Delta t = 120$ sec, for AAPL, AMZN, MSFT and CSCO stocks. The effect of clustering for MSFT and CSCO stocks is clearly visible.

Here $S(x)$ is the CDF of the data for the observations with value at least x_{min}, and $P(x)$ is the CDF of the power-law model that best fits the data in the region $x \geq x_{min}$. The estimate $\hat{x}_{min}$ is the value of x_{min} that minimizes D.

Results:

We compute the hypercumulants for our NASDAQ stocks at different time scales. In trade time we use for $\Lambda_q(n)$ the following set of values for n: 1, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192. In continuous time we use for $\Lambda_q(\Delta t)$ the following set of values for Δt measured in seconds: 1, 2, 5, 15, 30, 60, 120, 240, 480, 960, 1800, 3600 and 7200. Since the mean number of transactions in one day ranges from $12,000$ to $21,000$, the highest values of n represent a significant fraction of the entire day. In continuous time we investigate from the scale of one second to two hours.

We can observe from figs. 3.9, 3.10 and 3.11, 3.12 that the behavior of the hypercumulants for the two classes of stocks, i.e. large and small relative tick size, is different at small time scales but that they seem to converge at higher time scales. In these figures we compute the corresponding value of Λ_q for the standard Gaussian distribution in order to control the convergence of distributions of price returns as a function of the time scale of aggregation n or Δt. We show the dependence of the hypercumulants from the aggregation time scale in figs. 3.11, 3.12 for some fixed values of q. When we have a large relative tick stock, the convergence of normalized returns g to a Gaussian behavior is slower with respect to returns computed in presence of a small relative tick stock. A possible motivation could be the presence of clustering of returns for large tick size stocks, i.e. we have a distortion of distribution that is almost absent for the case of small tick size stocks. We make this hypothesis on the

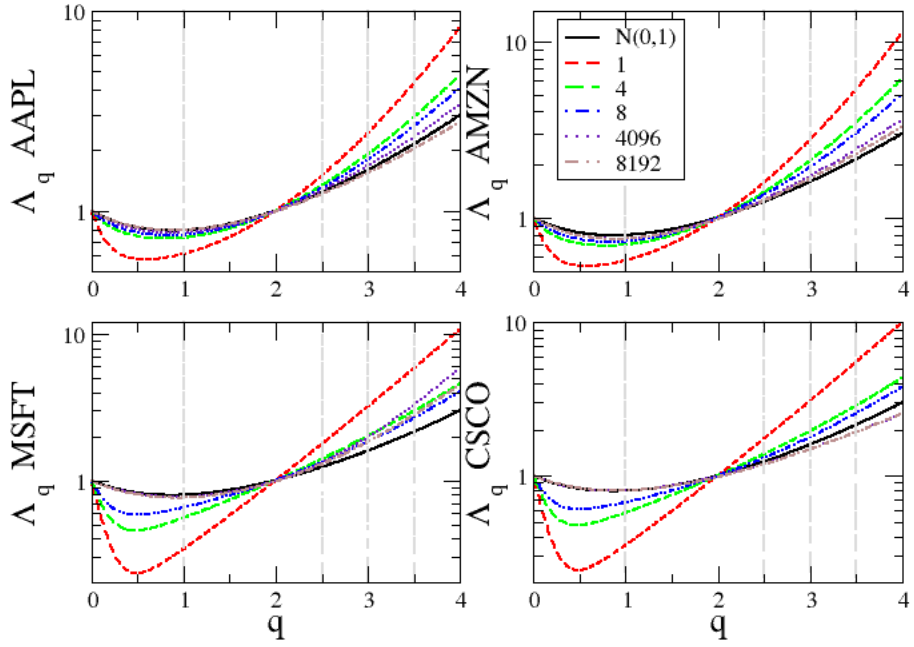

Figure 3.9: Linear-log plot of the scaling of hypercumulants Λ_q of normalized returns g in trade time. The different style of lines indicate the different time scales Δt. The vertical lines indicate the values of q, i.e. $q = 1, 3$, for which we illustrate the time scale dependence $\Lambda_q(n)$ in fig. 3.11.

basis of the computed histograms of distributions of returns. Here we can observe that the discretization effects starts to disappear after $n = 32$ for a small tick size stock, instead this threshold is higher for large tick size stocks (where we have also the presence of clustering), i.e. we find a value of $n = 512 \div 1024$. If we observe the returns process in continuous time, the discretization effect disappears after a time scale around $\Delta t \simeq 30$ secs. for small tick stocks, instead for a large tick stock it starts to disappear after around 16 mins.

We observe that independently of the relative tick size, the convergence toward a Gaussian behavior in continuous time is slower than in trade time. This observation may be explained by the subordination hypothesis. The original idea dates back to a paper by Mandelbrot [182, 181] that was later developed by Clark [63]. Mandelbrot and Taylor proposed that prices could be modeled as a subordinated random process $Y(t) = X(\tau(t))$, where Y is the random process generating returns, X is a Brownian motion and $\tau(t)$ is a stochastic time clock whose increments are i.i.d. and uncorrelated with the process X. Clark hypothesized that the time clock $\tau(t)$ is the cumulative trading volume in time t, but more recent works indicated that the number of transactions, i.e. trade time, is more important than their size [20]. Gillemot et al.

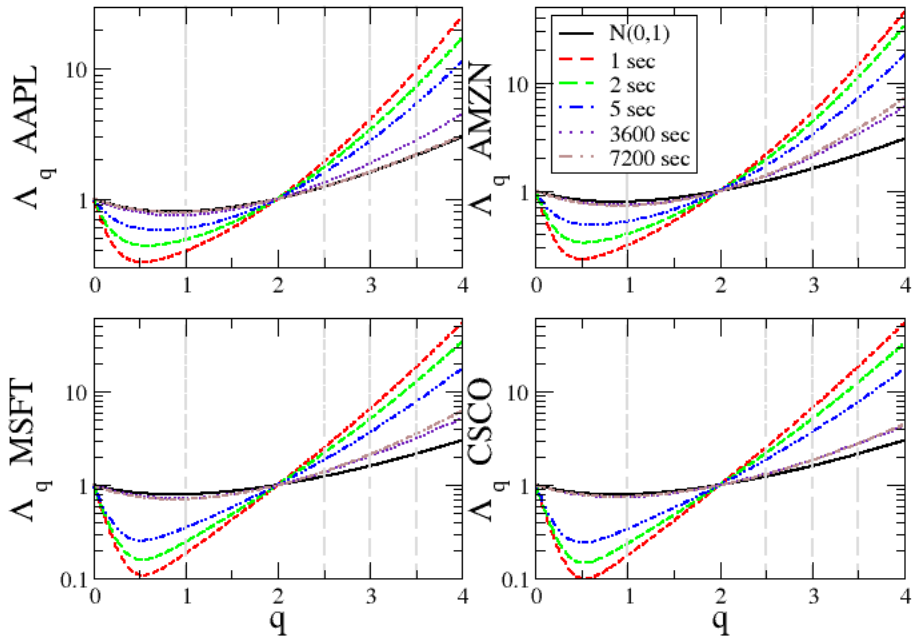

Figure 3.10: Linear-log plot of the scaling of hypercumulants Λ_q of normalized returns g in continuous time. The different style of lines indicate the different time scales Δt. The vertical lines indicate the values of q, i.e. $q = 1, 3$, for which we illustrate the time scale dependence $\Lambda_q(n)$ in fig. 3.12.

[113] and La Spada et al. [154] showed that the role of the subordination hypothesis in fat tails of returns is strongly dependent on the tick size. From our point of view it is important to stress that the stochastic clock $\tau(t)$ could modify the moments of distributions for the increments ΔX and ΔY. For example Clark [63] showed that if X is a Gaussian stochastic process with stationary independent increments, and $\tau(t)$ has stationary independent positive increments with finite second moment which are independent from X, then the kurtosis of the increments of $X(\tau(t))$ is an increasing function of the variance of the increments of $\tau(t)$. Our hypothesis is that a similar effect could be the motivation of the distortion of the value of the hypercumulants for the same value of time scale aggregation, i.e. the distribution of the sum of $n = 100$ identically distributed values of 1 transaction returns is not the same distribution that we obtain summing 100 identically distributed values of 1 sec. returns.

Let discuss now the results of the estimation of the tail exponent α of the distribution of returns at different time scales of aggregation. Fig. 3.13 shows the estimated values of α with the error bars for aggregation in transaction time (left panel) and in real time (right panel). For small values of aggregation (in real or transaction time) a clear difference appears between large and small tick size stocks. The former type of

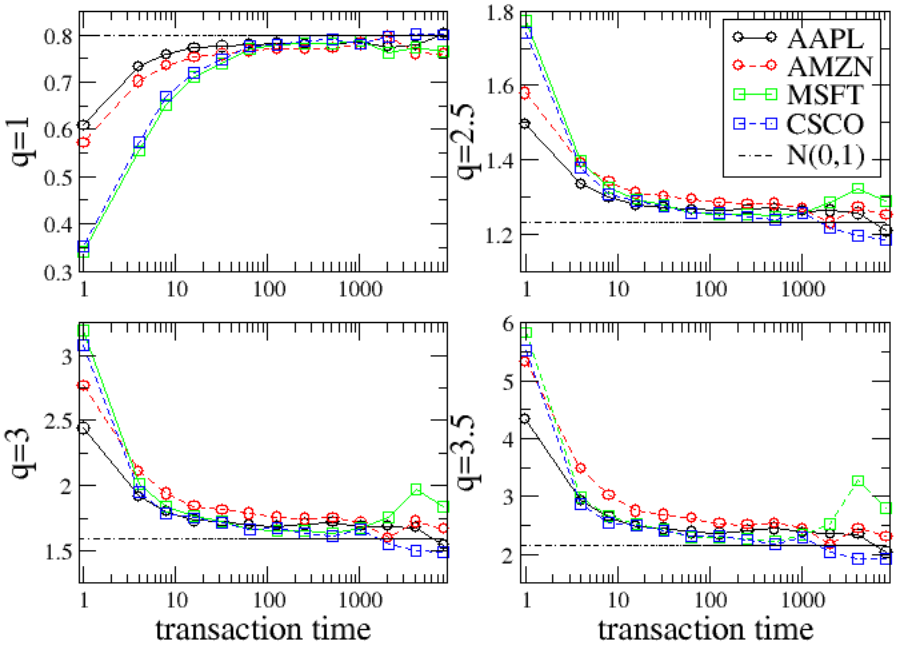

Figure 3.11: Log-linear plot of the scaling of hypercumulants $\Lambda_{q=1,2.5,3,3.5}(n)$, computed in trade time, for stocks: AAPL, AMZN, MSFT and CSCO. We observe a different speed of convergence to a Gaussian behavior between small and large tick size stocks.

stocks displays a large estimated value of α, while for the latter class the exponent is already quite small. When the aggregation scale increases, the estimated tail exponent for large tick size stocks rapidly decays and around $n \simeq 30$ or $\Delta t \simeq 30$ secs their behavior becomes indistinguishable from that of small tick stocks. This suggests that the same underlying and latent price process characterizes large and small tick size stocks, but for the former class the large tick size hides the process, at least until the crossover time scale.

After this crossing, the estimated exponent monotonically decreases (at least until the maximal investigated time scale). It is worth noticing that at the largest time scales of aggregation the estimated exponent, both in real and in transaction time, is smaller than 2, indicating a Levy-like regime. Before commenting this result it is important to stress that the values shown in fig. 3.13 are *estimated* exponents, i.e. there is no guarantee that the distibutions have a true power law tail. The Clauset et al. [64] algorithm gives the best estimate of α and x_{min} *assuming* that the tail is power law. Moreover as noticed by Clauset when the sample is small, the method can give incorrect estimations. This is the case of the last two points of fig. 3.13, where the sample size is around 100 data points. These two considerations highlight the

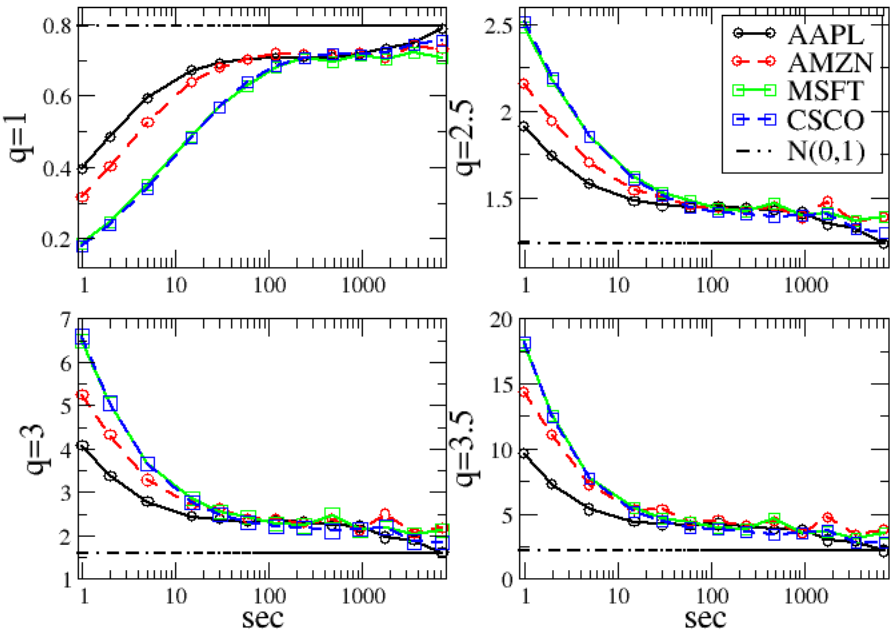

Figure 3.12: Log-linear plot of the scaling of hypercumulants $\Lambda_{q=1,2.5,3,3.5}(\Delta t)$, computed in continuous time, for stocks: AAPL, AMZN, MSFT and CSCO. We observe a different speed of convergence to a Gaussian behavior between small and large tick size stocks.

problems that could arise when estimating the tail exponent of a distribution which has a discrete (and finite) support. In this case the tail is obviously not power law, but the method gives in any case an estimated value. In fact numerical simulations of iid models with finite support and thus finite variance (the model i.i.d. discussed in the next section) display a similar behavior of the estimated tail exponent, including a value smaller than 2 for large aggregation. This is clearly a misestimation result.

In conclusion, the analysis of the tail exponent shows two regimes, one in which large and small tick size stocks show a markedly different behavior and one where tick-to-price ratio does not play any role. Moreover numerical simulations and empirical analyses suggest to be very cautious when estimating the tail exponent of a distribution that is either defined on a discrete support (as for price changes) or has an hidden discretization (as for log-returns). Arbitrary small values of the exponent could be (mis)estimated as a result of an improper use of statistical methods.

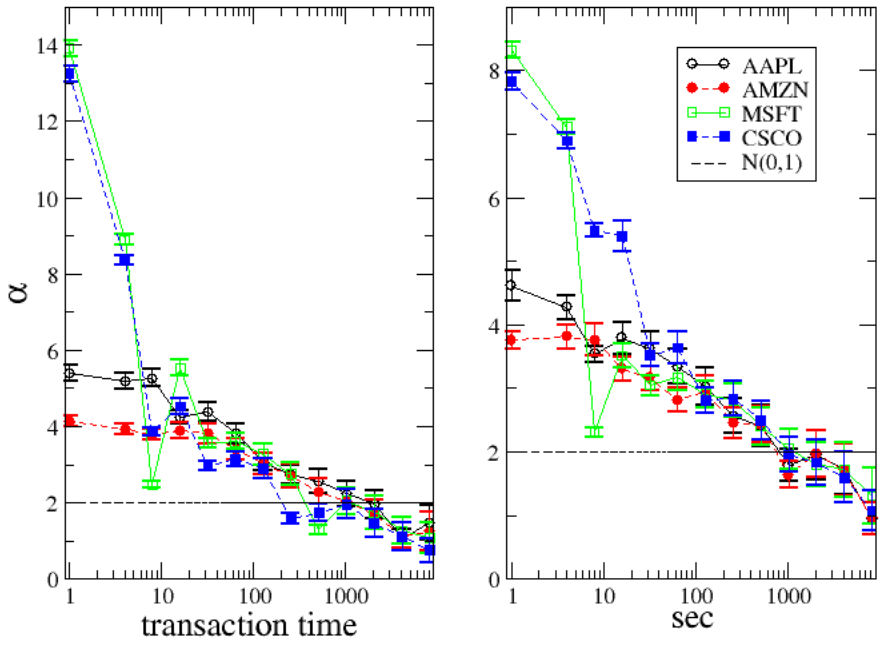

Figure 3.13: Log-linear plot of the scaling of asymptotic tail exponent α for distribution of normalized returns g as a function of trade time n and continuous time Δt for stocks: AAPL, AMZN, MSFT and CSCO. The straight dashed line $\alpha = 2$ is the upper bound of Lévy behavior, i.e. $0 < \alpha < 2$. The horizontal dotted line is the upper bound of the index for which we computed Λ_q, $q \in [1, 4]$.

3.3 Tick size and bid-ask spread statistical properties

The distributions of the bid-ask spread are heavily influenced by the value of the relative tick size, as we can observe in fig. 3.14. The main difference is that for large tick stocks the spread is almost always one tick. Instead, for small tick stocks the discrete distribution of the spread is more dispersed.

Another empirical property is that the autocorrelation function of the spread series decay very slowly, i.e an indication of long-range dependence. This behavior is well known in literature [123] and is shared by the two class of stocks.

There are several ways to characterize long-memory processes. One way to define long memory is via the asymptotic behavior of the autocorrelation function. A time series is said to exhibit long memory if

$$\lim_{\tau \to \infty} \rho(\tau) \,/\, \left(c\,\tau^{-\alpha}\right) = 1, \tag{3.7}$$

where $\alpha \in (0,1)$, $c > 0$, and $\rho(\tau)$ denotes the τ th-order autocorrelation. An immediate consequence is that autocorrelations are not absolutely summable. The exponent α describes the strength of the long memory, i.e. the smaller the value of α, the stronger is the long-range autocorrelations [249]. Because of the slow decay of the autocorrelation function in a long memory process, present values of the series can have significant effect on its values in the distant future. Long range dependence is often found in financial market data [27, 164].

A key difficulty when using equation 3.7 to asses whether a given series has long memory is that it deals only with asymptotic behavior. Also, the values of the autocorrelation function can be small, making estimation very difficult. Therefore, direct estimation of α from the autocorrelation function often produces very poor results [164]. Here, we use the *log-periodogram regression* (GPH test) [111] to test the presence of long range behavior of the spread series. This technique, proposed by Geweke and Porter-Hudak (GPH), is based on observations of the slope of the spectral density function of a fractionally integrated series around the angular frequency $\omega = 0$. It gives an estimate of the fractional differencing parameter d [27] that is connected to the exponent α simply by $\alpha = 1 - 2\,d$. There is evidence of long memory if the parameter d is significant and in the range $(0, 0.5)$.

The estimation procedure begins with calculating the periodogram, which is a sample analogue of the spectral density. For a vector of observations $\{x_1, \cdots, x_N\}$ the periodogram is defined as

$$I_N(\omega_k) = \frac{1}{N}\left|\sum_{t=1}^{N} x_t e^{-2\pi i(t-1)\omega_k}\right|^2, \tag{3.8}$$

where $\omega_k = k/N$, $k = 1, \cdots, [N/2]$ and $[x]$ denotes the largest integer less than or equal to x. The next and final step is to run a simple linear regression

$$\log\{I_N(\omega_k)\} = a - d\log\left\{4\sin^2(\omega_k/2)\right\} + \epsilon_k, \tag{3.9}$$

at lower Fourier frequencies ω_k, $k = 1, \cdots, K \leq [N/2]$. The least squares estimate of the slope yields the differencing parameter d. A major issue on the application of this method is the choice of K. Geweke and Porter-Hudak, as well as a number of other authors, recommend choosing K such that $K = \left[\sqrt{N}\right]$. We follow this recommendation.

From fig. 3.15 we can observe the sample autocorrelation for large and small tick stocks. In table 3.2 we report the estimates of the parameter d. For all stocks the estimate is significantly different from zero. However the large tick stocks show a lower value for the parameter d with respect to small tick stocks. To summarize, the main difference that we find is that the memory of the spread process for large tick stocks is lower than the one of small tick stocks.

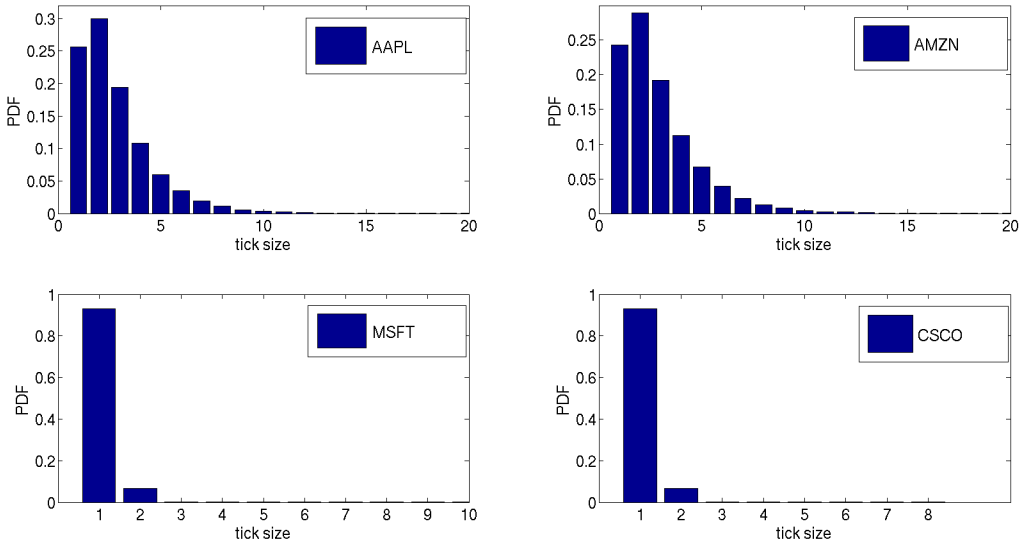

Figure 3.14: Distributions of bid-ask spread for small and large tick stocks.

	AAPL	AMZN	MSFT	CSCO
mean	2.78	2.93	1.07	1.07
sdev	1.94	2.15	0.27	0.27
max	52	65	14	8
min	1	1	1	1
median	2	2	1	1

Table 3.1: Summary statistics for spread data, measured in units of one tick size.

d	AAPL	AMZN	MSFT	CSCO
Estimate	0.4427	0.4554	0.1945	0.1854
t-statistic	21.55	19.25	8.28	7.44
p-value	0.000	0.000	0.000	0.000

Table 3.2: GPH test of the spread series.

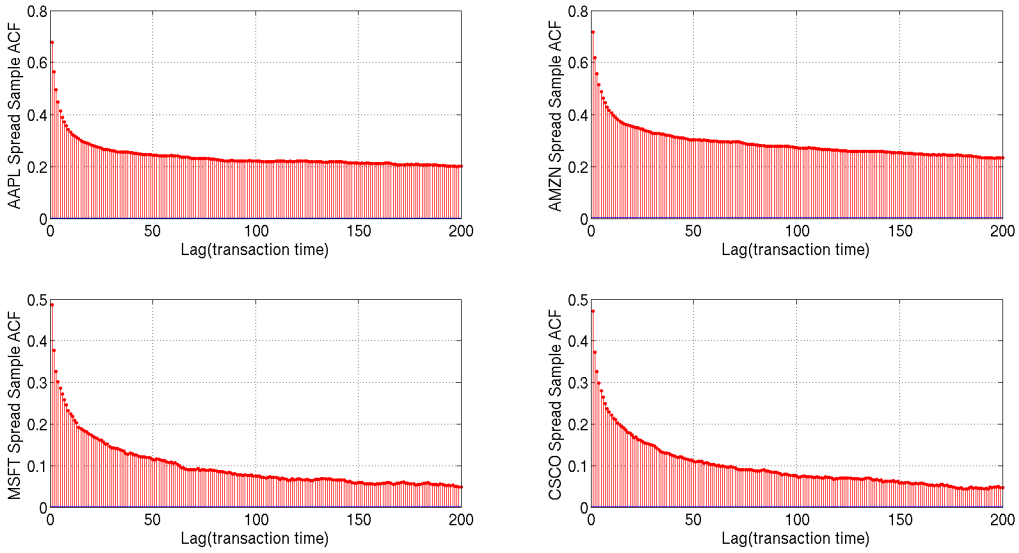

Figure 3.15: Sample autocorrelations for small and large tick stocks. Lags are given in transaction time.

Chapter 4

Modeling the coupled price-spread dynamics of large tick stocks

Large tick assets, i.e. assets where one tick movement is a significant fraction of the price and bid-ask spread is almost always equal to one tick, display a dynamics in which price changes and spread are strongly coupled. We present an approach based on the hidden Markov model, also known in econometrics as Markov switching model, for the dynamics of price changes, where the latent Markov process is described by the transitions between spreads. We then use a finite Markov mixture of logit regressions on past squared price changes to describe temporal dependencies in the dynamics of price changes. The model can thus be seen as a double chain Markov model. We show that the model describes the shape of price changes distribution at different time scales, volatility clustering, and the anomalous decrease of kurtosis. We calibrate our models on Nasdaq stocks and we show that this model reproduces remarkably well the statistical properties of real data.

4.1 Introduction

In this study we are interested in modeling the dynamics of large tick assets at ultra high frequency [10] and taking explicitely into account the discreteness of prices. We propose zero-intelligence models [82] for prices and spread dynamics based on the hidden Markov and double chain Markov processes [103, 55, 256], [30]. Hidden Markov models are widely used in the field of complex systems: condensed matter [141], DNA segmentation [234], econophysics [91], neuro-science [241] and biochemistry [151], [158]. In the econometric literature hidden Markov processes [214] are also known as Markov switching models [128], in this work we use indifferently the two terms. A modeling approach for the limit order book's dynamics based on Markov or on matrix multiplicative processes was already used in the past [70, 229]. However, in this models the Markov process describes the arrivals of market orders, limit orders and order cancellations [130] and not directly price and bid-ask spread dynamics. D'Amico and Petroni [79, 78, 80], instead, have proposed a model based on semi-Markov chains for studying directly high frequency financial price changes. This

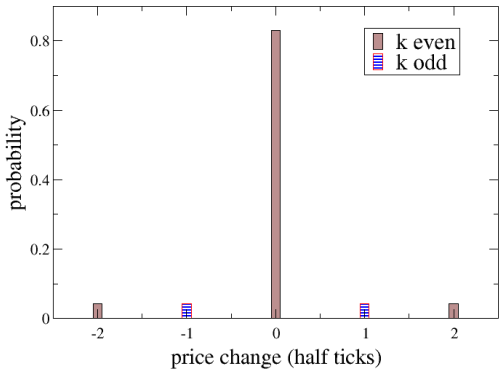

Figure 4.1: (Color online) Distribution of mid-price change between two transactions, $\Delta p\left(t, \Delta t = 1\right) = p_m\left(t+1\right) - p_m\left(t\right)$. The investigated stock is Microsoft.

model is able to reproduce the autocorrelation of squared price changes observed at 1 min frequency.

Here, we introduce a class of models describing the coupled dynamics of price changes and spread for large tick assets in transaction time[5]. Therefore, these models are defined in a discrete state space [185, 163] and the time evolution is described in discrete time. Our purpose is to model price dynamics in order to reproduce statistical properties of mid-price dynamics at different time scales and stylized facts like volatility clustering [69]. Notice that, rather than considering a non observable efficient price [42] and describing the data as the effect of the round off error due to tick size [155], we directly model the observable quantities, such as spread and mid-price, by using a time series approach [128].

The motivation of our work comes from two interesting empirical observations. Let us consider first the unconditional distribution of mid-price changes at different time scales. In Fig. 4.1 we show the histogram of mid-price change of MSFT at the finest transaction time scale, i.e. the change between two transactions. It is clear that most of the times the price does not change, while sometimes it changes by one or two half ticks. When we aggregate the price changes on a longer time scale, for example 128 transactions (see the panel of Fig. 4.2), a non trivial distribution emerges, namely a distribution where odd values of price changes are systematically less populated than even values. It is important to notice that if we assume that price changes of individual trades are independent and identically distributed[7], we would never be able to reproduce an histogram like the one shown in the panel of Fig. 4.2. In fact in this case the histogram would be, as expected, bell shaped. In a recent

[5]Hereafter we define the transaction/trade time t as an integer counter of events defined by the execution of a market order. Note that if a market order is executed against several limit orders, our clock advances only by one unit. We choose to use the variable t, instead of the variable i used in section 2.2.1, in order to discriminate the time variable respect to the other integer variables. The integer variables are heavily used in a Markov modeling context.

[7]For example, if we randomize our sample of transaction to transaction mid-price changes

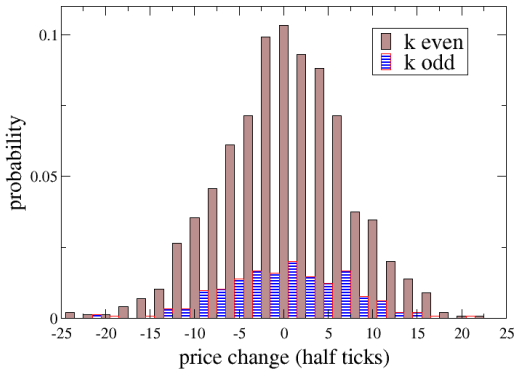

Figure 4.2: (Color online) Mid-price change distribution aggregated at 128 transactions, $\Delta p\left(t, \Delta t = 128\right) = p_m\left(t + 128\right) - p_m\left(t\right)$. The investigated stock is Microsoft.

study [253] Zaccaria et al. have observed an asymmetry between even and odd values of the spread and, in the context of the order book model of Ref. [75], it has been related to the rate of limit orders. As we will show below, this spread asymmetry and the coupling between spread and price changes can be used to explain the mid-price change distribution observed in Fig. 4.2. In fact, if odd spread values are much more probable than even values, than even price changes are much more probable than odd price changes (see Figure 4.4 below for an intuition).

The second observation concerns the properties of volatility of the price change process. Figure 4.3 shows the autocorrelation function of squared price changes of MSFT in transaction time. Square price changes can be seen here as a simple proxy of volatility. First of all notice that the autocorrelation is negative for small lags. It then reaches a maximum around 10 trades and then it decays very slowly to zero. We observe that between 10 and more than 500 trades, the decay of the autocorrelation function is well described by a power law function, $corr\left(\Delta p^2\left(t\right), \Delta p^2\left(t+\tau\right)\right) \sim \tau^{-\gamma}$, and the estimated exponent $\gamma \simeq 0.3$ is similar to the one observed at lower frequency and by sampling price changes in real time rather than transaction time[8]. We conclude therefore that very persistent volatility clustering and possibly long range volatility is observed also at the transaction to transaction time scale.

The purpose of this study is to develop a discrete time series model that is able to explain and reproduce simultaneously these two empirical observations, namely the change of the distribution of price changes at different time scales and the shape of the volatility autocorrelation describing its time clustering. The key intuition behind our modeling approach is that for large tick assets the dynamics of mid-price and of spread are intimately related and that the process of price changes is conditioned

[8]It is worth noticing that in general the round-off error severely reduces the correlation properties of a stochastic process, even if the Hurst exponent of a long memory process is preserved [155]. Therefore the autocorrelation function shown in Fig. 4.3 is a strong underestimation of the transaction to transaction volatility clustering of the unobservable efficient price.

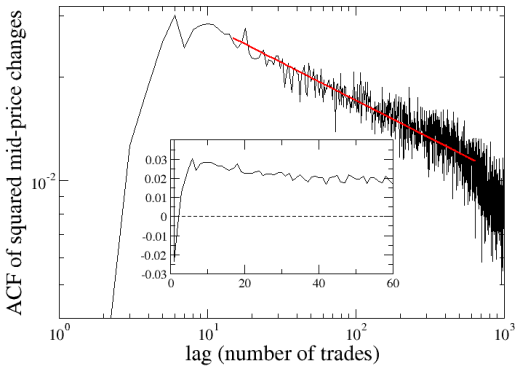

Figure 4.3: (Color online) Sample autocorrelation function of transaction to transaction squared mid-price changes for Microsoft, reported in $\log_{10}$-$\log_{10}$ scales. The plot is in log-log scale and the red dashed line is a best fit of the autocorrelation function in the considered region. The estimated exponent is $\gamma = 0.301$. The inset shows the behavior for small values of the lag.

to the spread process. The conditioning rule describes the connection between the stochastic motion of mid-price and spread on the price grid.

More specifically, for large tick assets the spread typically assumes only few values. For example, for MSFT spread size is observed to be 1 or 2 ticks almost always. The discreteness of mid-price dynamics can be connected to the spread dynamics if we observe that, when the spread is constant in time, price changes can assume only even values in units of half tick. Instead when the spread changes, price changes can display only odd values. Figure 4.4 shows the constrained relation between the two processes. The dynamics of price changes is thus linked to dynamics of spread transitions. This relation leads us to design models in which the price change process depends on the transition between two subsequent spread states, distinguishing the case in which the spread remains constant and the case when it changes. From a methodological point of view we obtain this by defining a variable of state that describes the spread transition. We use a hidden Markov, or Markov switching, model [214, 128] for price changes, in which the spread transition is described by a Markov chain that defines different regimes for the price change process.

The Markov switching approach is able to describe the change in shape of the distribution of mid-price change (Figs. 4.1, 4.2), but not the persistence of volatility. To this end, we propose a more sophisticated model by allowing the price change process to be an autoregressive process in which regressors are the past value of squared price changes [30, 31, 32]. We show how to calibrate the models on real data and we tested them on the large tick assets MSFT and CSCO, traded at NASDAQ market in the period July-August 2009. We show that the full model reproduces very well the empirical data.

The study is organized as follows. In Section 4.2 we review the main applications of Markov switching modeling in the econometrics field. In Section 4.3 we present our

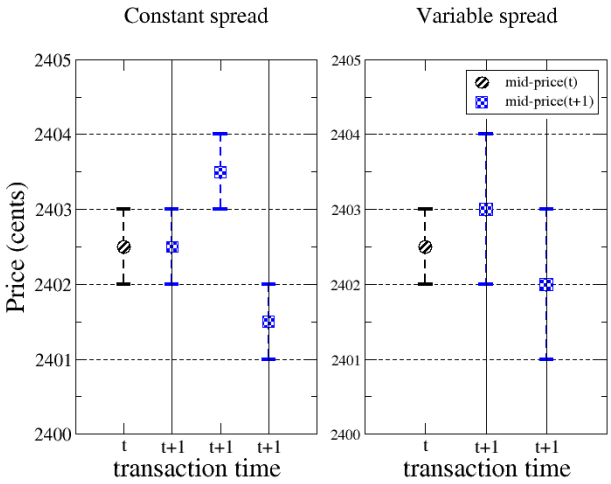

Figure 4.4: (Color online) Coupling of spread and price changes for large tick assets. On the left we show the three possible transitions when $s(t) = s(t+1) = 1$. In this case the possible price changes are $\Delta p(t) \in (-2, 0, 2)$ (measured in $1/2$ tick size). On the right we show the two possible transitions when $s(t) = 1$ and $s(t+1) = 2$. In this case the possible values of price changes are $\Delta p(t) \in (-1, 1)$.

modeling approach. In Section 4.4 we present our data for the MSFT stock and we describe the observed stylized facts of price dynamics. In Section 4.5 we describe the calibration of the models on real data and we discuss how well the different models reproduce the stylized facts. Finally, in Section 4.6 we draw some conclusions and we discuss future works.

4.2 Review of Markov switching models in econometrics

Markov switching models (MS models) have become increasingly popular in econometric studies of industrial production, interest rates, stock prices and unemployment rates [127, 103]. They are also known as hidden Markov models (HMM) [214, 53], used for example in speech recognition and DNA analysis. In these models the distribution that generates an observation depends on the states of an underlying and unobserved Markov process. They are flexible general purpose models for univariate and multivariate time series, especially for discrete-valued series, including categorical variables and series of counts [256]. Markov switching models belong to the general class of finite mixture models [103]. Econometricians' initial interest in this class of distributions was based on their ability to flexibly approximate general classes of density functions and generate a wider range of values for the skewness and kurtosis than is obtainable by using a single distribution. Along these lines Granger and Orr [120] and Clark [63] considered time-independent mixtures of normal dis-

tributions as a means of modeling non-normally distributed data. These models, however, did not capture the time dependence in the conditional variance found in many economic time series, as evidenced by the vast literature on ARCH models that started with Engle [93, ?]. By allowing the mixing probabilities to display time dependence, Markov switching models can be seen as a natural generalization of the original time-independent mixture of normals model. Timmermann [235] has shown that the mixing property enables them to generate a wide range of coefficients of skewness, kurtosis and serial correlation even when based on a very small number of underlying states. Regime switches in economic time series can be parsimoniously represented by Markov switching models by letting the mean, variance, and possibly the dynamics of the series depend on the realization of a finite number of discrete states.

The basic MS model is:

$$y(t) = \mu_{S(t)} + \sigma_{S(t)}\epsilon(t), \tag{4.1}$$

where $S(t) = 1, 2, \cdots, k$ denotes the unobserved state indicator which follows an ergodic k-state Markov process and $\epsilon(t)$ is a zero-mean random variable which is i.i.d. over time. Another relevant model is the Markov switching autoregressive model (MSAR(q)) of order q that allows for state-independent autoregressive dynamics:

$$y(t) = \mu_{S(t)} + \sum_{j=1}^{q} \phi_j \left(y(t-j) - \mu_{S(t-j)} \right) + \sigma_{S(t)}\epsilon(t). \tag{4.2}$$

It became popular in econometrics for analyzing economic time series such as the GDP data through the work of Hamilton [127]. In its most general form the MSAR model allows that the autoregressive coefficients are also affected by $S(t)$ [235]:

$$\begin{aligned} y(t) &= \mu_{S(t)} + \sum_{j=1}^{q} \phi_{j,S(t-j)} \left(y(t-j) - \mu_{S(t-j)} \right) + \\ &\quad + \sigma_{S(t)}\epsilon(t). \end{aligned} \tag{4.3}$$

There is a key difference with respect to ARCH models, which is another type of time-dependent mixture processes. While Markov switching models mix a finite number of states with different mean and volatility parameters based on an exogenous state process, ARCH models mix distributions with volatility parameters drawn from an infinite set of states driven by lagged innovations to the series.

The above described models can be used when the variable $y(t)$ under investigation is continuous. In our case the observed price difference is instead a discrete variable. Therefore the models for continuous variables presented above cannot be used in our problem. We propose to model the coupled dynamics of spreads and price differences in the setting defined by the double chain Markov models (DCMM) [30, 31]. This is the natural extension of HMM models in order to allow the hidden Markov process to select one of a finite number of Markov chains to drive the observed

process at each time point. If a time series can be decomposed into a finite mixture of Markov chains, then the DCMM can be applied to describe the switching process between these chains. DCMM belongs to the family of Markov chains in random environments [66, 67].

In discrete time, DCMM describes the joint dynamics of two random variables: $x(t)$, whose state at time t is unknown for an observer external to the process, and $y(t)$, which is observable. The model is described by the following elements:

- A set of hidden states, $\mathcal{S}(x) = \{1, \cdots, N_x\}$.
- A set of possible outputs, $\mathcal{S}(y) = \{1, \cdots, N_y\}$.
- The probability distribution of the first hidden state, $\boldsymbol{\pi_0} = \{\pi_{0,1}, \cdots, \pi_{0,N_x}\}$.
- A transition matrix between hidden states, $M = \{m_{ij}\},\ i,\ j \in \mathcal{S}(x)$.
- A set of transition matrices between successive outputs of $y(t)$ given a particular state of $x(t)$, $V_{x(t)=k,ij},\ i,\ j \in \mathcal{S}(y),\ k \in \mathcal{S}(x)$.

There are three different estimation problems: the estimation of the probability of a sequence of observations $y(0), \cdots, y(T)$ given a model; the estimation of parameters $\boldsymbol{\pi_0},\ M,\ V_k$ given a sequence of observations; the estimation of the optimal sequence of hidden states given a model and a sequence of outputs.

Our limit order book data instead allow us to see directly the process that defines the hidden Markov process, i.e. the spread process. In this way we can estimate directly the matrices M and V_k by a simple maximum likelihood approach, without using the Expectation Maximization (EM) algorithm and the Viterbi algorithm, that are usually used when the hidden process is not observable [30, 31]. We use the stationary probability distribution for the process $x(t)$ as initial probability distribution $\boldsymbol{\pi_0}$ in order to perform our calculations and simulations. We use the DCMM model as a mathematical framework for spread and price differences processes without treating spread process as an hidden process.

Among the few financial applications of the DCMM model we mention Ref.s [100, 89]. In the former paper, authors studied the credit rating dynamics of a portfolio of financial companies, where the unobserved hidden process is the state of the broader economy. In Eisenkopf [89] instead the author considered a problem in which a credit rating process is influenced by unobserved hidden risk situations. To the best of our knowledge our paper is the first application of DCMM to the field of market microstructure and high frequency financial data.

4.3 Markov models for the coupled dynamics of spread and price changes

In this section we present the two classes of models describing the process of price changes $\Delta p(t, \Delta t) = p_m(t + \Delta t) - p_m(t)$ at time scale Δt, where we define the mid-price as $p_m(t) = (p_{ASK}(t) + p_{BID}(t))/2$ and we choose to measure Δp in units of half

tick size. In our models, price change process follows different time series processes conditioned on the dynamics of transitions of the spread $s(t) = p_{ASK}(t) - p_{BID}(t)$. Hereafter we will use the notation $\Delta p(t) = r(t, \Delta t = 1)$. The spread variable s is measured in units of 1 tick size, so we have $\Delta p(t, \Delta t) \in \mathbb{Z}$ and $s(t) \in \mathbb{N}$. The time variable $t \in \mathbb{N}$ is the transaction time. The first class of models is a Markov switching model and it is able to describe the empirically observed change of the distribution of price change, but it is unable to describe the volatility clustering. The intuition behind this approach relies on the empirical observation that the value of the spread is almost always one tick. This should imply a distribution of price change populated only by even values at each time scale. In order to model the presence of a small population of odd values at low frequencies, we introduce a simple Bernoulli or Markov dynamics for a two-state spread process. The second more complex class of models, a double chain Markov model, uses the past price dynamics to determine the probabilities and it is able to describe all the investigated features of price.

4.3.1 Markov switching models

Spread process. It is well known that spread process is autocorrelated in time [206, 123, 208]. Moreover, in section 3.3 we have verified that for the stocks under study this property holds. We model the spread $s(t)$ as a stationary Markov(1) [125] process[1]:

$$\begin{aligned} P \quad & (s(t) = j | s(t-1) = i, s(t-2) = k, \cdots) = \\ = \quad & P(s(t) = j | s(t-1) = i) = p_{ij}, \end{aligned} \tag{4.4}$$

where $i, j \in \mathbb{N}$ are spread values. As mentioned, we limit the set of spread values to $s \in \{1, 2\}$, because we want to describe the case of large tick assets. We also assume that the process $s(t)$ is not affected by the price change process $\Delta p(t)$[2]. The spread process is described by the transition matrix $B \in M_{2,2}(\mathbb{R})$:

$$B = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix},$$

where the normalization is given by $\sum_{j=1}^{2} p_{ij} = 1$. The vector of stationary probabilities is the eigenvector $\boldsymbol{\pi}$ of B' relative to eigenvalue 1, which is:

$$B'\boldsymbol{\pi} = \boldsymbol{\pi}, \qquad \boldsymbol{\pi} = \begin{pmatrix} (1 - p_{22}) / (2 - p_{11} - p_{22}) \\ (1 - p_{11}) / (2 - p_{11} - p_{22}) \end{pmatrix}, \tag{4.5}$$

where B' denotes the transpose of the matrix B. This vector represents the unconditional probabilities of $s(t)$, so $\pi_k = P(s(t) = k)$ with $k = 1, 2$.

[1] We have tried other specifications of the spread process, such as for example a long memory process, but this does not change significantly our results. We clarify this point in section 4.9.1.

[2] This simplifying assumption does not imply that the switching models cannot describe implicitly a conditioning of the spread to price changes. It can be shown that our models succeed to reproduce the empirical conditional distributions $P(S(t+1) | \Delta p(t))$. This is reported in section 4.9.2

Starting from the $s(t)$ process, it is useful to define a new stationary Markov(1) process $x(t)$ that describes the stochastic dynamics of transitions between states $s(t)$ and $s(t+1)$ as:

$$\begin{aligned} x(t) &= 1 \quad \text{if} \quad s(t) = 1 \to s(t+1) = 1, \\ x(t) &= 2 \quad \text{if} \quad s(t) = 1 \to s(t+1) = 2, \\ x(t) &= 3 \quad \text{if} \quad s(t) = 2 \to s(t+1) = 1, \\ x(t) &= 4 \quad \text{if} \quad s(t) = 2 \to s(t+1) = 2. \end{aligned} \tag{4.6}$$

This Markov process is characterized by a new transition matrix $M \in M_{4,4}(\mathbb{R})$:

$$M = \begin{pmatrix} p_{11} & p_{12} & 0 & 0 \\ 0 & 0 & p_{21} & p_{22} \\ p_{11} & p_{12} & 0 & 0 \\ 0 & 0 & p_{21} & p_{22} \end{pmatrix},$$

where some transitions $x(t) \to x(t+1)$ are forbidden because do not correspond to any allowed spread sequence $s(t) \to s(t+1) \to s(t+2)$. The stationary vector of M is given by:

$$M'\boldsymbol{\lambda} = \boldsymbol{\lambda}, \ \boldsymbol{\lambda} = \begin{pmatrix} (p_{21}p_{11}) / (1 - p_{11} + p_{21}) \\ p_{21}(1 - p_{11}) / (1 - p_{11} + p_{21}) \\ p_{21}(1 - p_{11}) / (1 - p_{11} + p_{21}) \\ (1 - p_{21})(1 - p_{11}) / (1 - p_{11} + p_{21}) \end{pmatrix}. \tag{4.7}$$

A limiting case is when the spread process $s(t)$ is described by a Bernoulli process. In this case we set $P(s(t) = 1) = p$. Although $s(t)$ is an i.i.d. process, the spread transition process $x_B(t)$ is a Markov process defined by:

$$M_B = \begin{pmatrix} p & (1-p) & 0 & 0 \\ 0 & 0 & p & (1-p) \\ p & (1-p) & 0 & 0 \\ 0 & 0 & p & (1-p) \end{pmatrix}, \quad \boldsymbol{\lambda}_B = \begin{pmatrix} p^2 \\ p(1-p) \\ p(1-p) \\ (1-p)^2 \end{pmatrix}.$$

In the general case, the process $x(t)$ is defined by two parameters p_{11}, p_{21} (which are reduced to p in Bernoulli case) that we can estimate from spread data.

Mid-price process. We can now define a Markov switching process for price changes $\Delta p(t)$ which is conditioned to the process $x(t)$, i.e. to the spread transitions. Returns are measured in half ticks and we limit the set of possible values to $\Delta p(t) \in \{-2, -1, 0, 1, 2\}$, as observed in our sample (see Figure 4.1). The discreteness

of the price grid imposes the mechanical constraints:

$$\begin{aligned} x(t) = 1 \quad &\longrightarrow \quad \Delta p(t) \in \{-2, 0, 2\}, \\ x(t) = 2 \quad &\longrightarrow \quad \Delta p(t) \in \{-1, 1\}, \\ x(t) = 3 \quad &\longrightarrow \quad \Delta p(t) \in \{-1, 1\}, \\ x(t) = 4 \quad &\longrightarrow \quad \Delta p(t) \in \{-2, 0, 2\}. \end{aligned} \tag{4.8}$$

The mapping between transitions $x(t)$ and allowed values of the mid-price changes $\Delta p(t)$ has been done by using the cases shown in Fig. 4.4. This assumption is grounded on the empirical observation that mid-price changes $|\Delta p(t)| > 2$ are extremely rare for large tick assets (see Section 4).

In the simplest model, we assume that the probability distribution of price changes between two transactions depends only on the spread transition between them. We can therefore define the following conditional probabilities defining the process of price changes:

$$\begin{aligned} P(\Delta p(t) = \pm 2 | x(t) = 1; \boldsymbol{\theta}) &= \theta_1, \\ P(\Delta p(t) = 0 | x(t) = 1; \boldsymbol{\theta}) &= 1 - 2\theta_1, \\ P(\Delta p(t) = \pm 1 | x(t) = 2; \boldsymbol{\theta}) &= 1/2, \\ P(\Delta p(t) = \pm 1 | x(t) = 3; \boldsymbol{\theta}) &= 1/2, \\ P(\Delta p(t) = \pm 2 | x(t) = 4; \boldsymbol{\theta}) &= \theta_4, \\ P(\Delta p(t) = 0 | x(t) = 4; \boldsymbol{\theta}) &= 1 - 2\theta_4. \end{aligned} \tag{4.9}$$

Notice that we have assumed symmetric distributions for price changes between positive and negative values and $\boldsymbol{\theta} = (\theta_1, \theta_4)'$ is the parameter vector that we can estimate from data. The parameter θ_1 (θ_4) describes the probability that mid-price changes when the spread remains constant at one (two) ticks.

The coupled model of spread and price change described here will be termed the MS model. When we consider the special case of spread described by a Bernoulli process we will refer to it as the MS_B model.

Properties of price price changes. Here we derive the moments and the autocorrelation functions $\zeta(\tau) \equiv corr(\Delta p(t), \Delta p(t+\tau))$ and $\rho(\tau) \equiv corr(\Delta p^2(t), \Delta p^2(t+\tau))$ under the MS model. The quantity $\zeta(\tau)$ is useful to study the statistical efficency of prices [43, 69], while $\rho(\tau)$ describes volatility clustering in transaction time.

We compute first the vectors of conditional first, second and fourth moments:

$$\begin{aligned} E[\Delta p(t) | x(t) = k] &= m_{1,k}, \\ E\left[\Delta p^2(t) | x(t) = k\right] &= m_{2,k}, \\ E\left[\Delta p^4(t) | x(t) = k\right] &= m_{4,k}, \end{aligned} \tag{4.10}$$

where $m_{j,k}$ indicates the $k-$th component of the vector $\boldsymbol{m}_j$. We have $\boldsymbol{m}_1 = \mathbf{0}$, $\boldsymbol{m}_2 = (8\theta_1, 1, 1, 8\theta_4)'$ and $\boldsymbol{m}_4 = (32\theta_1, 1, 1, 32\theta_4)'$. Then we compute unconditional

moments by using the stationary vector $\boldsymbol{\lambda}$ as:

$$
\begin{aligned}
&E\left[\Delta p\left(t\right)\right]=\sum_{k=1}^{4}E\left[\Delta p\left(t\right)|x\left(t\right)=k\right]P\left[x\left(t\right)=k\right]=\\
&=\boldsymbol{m}_{1}'\boldsymbol{\lambda},\\
&E\left[\Delta p^{2}\left(t\right)\right]=\sum_{k=1}^{4}E\left[\Delta p^{2}\left(t\right)|x\left(t\right)=k\right]P\left[x\left(t\right)=k\right]=\\
&=\boldsymbol{m}_{2}'\boldsymbol{\lambda},\\
&E\left[\Delta p^{4}\left(t\right)\right]=\sum_{k=1}^{4}E\left[\Delta p^{4}\left(t\right)|x\left(t\right)=k\right]P\left[x\left(t\right)=k\right]=\\
&=\boldsymbol{m}_{4}'\boldsymbol{\lambda},\\
&Var\left[\Delta p\left(t\right)\right]=\boldsymbol{m}_{2}'\boldsymbol{\lambda}-\left(\boldsymbol{m}_{1}'\boldsymbol{\lambda}\right)^{2},\\
&Var\left[\Delta p^{2}\left(t\right)\right]=\boldsymbol{m}_{4}'\boldsymbol{\lambda}-\left(\boldsymbol{m}_{2}'\boldsymbol{\lambda}\right)^{2},
\end{aligned}
\tag{4.11}
$$

In order to compute the linear autocorrelation function $\zeta(\tau)$ we need to compute $E\left[\Delta p\left(t\right)\Delta p\left(t+\tau\right)\right]$, by using conditional independence of $\Delta p\left(t\right)$ with respect to $x\left(t\right)$. We obtain:

$$
\begin{aligned}
&E\left[\Delta p\left(t\right)\Delta p\left(t+\tau\right)\right]=\\
&=\sum_{i=1}^{4}\sum_{j=1}^{4}\Big(E\left[\Delta p\left(t\right)\Delta p\left(t+\tau\right)|x\left(t\right)=i,x\left(t+\tau\right)=j\right]\\
&P\left[x\left(t\right)=i,x\left(t+\tau\right)=j\right]\Big)=\\
&=\sum_{i=1}^{4}\sum_{j=1}^{4}\Big(E\left[\Delta p\left(t\right)|x\left(t\right)=i\right]E\left[\Delta p\left(t+\tau\right)|x\left(t+\tau\right)=j\right]\\
&P\left[x\left(t\right)=i,x\left(t+\tau\right)=j\right]\Big)=\\
&=\sum_{i=1}^{4}\sum_{j=1}^{4}m_{1,i}m_{1,j}\lambda_{i}M_{ij}^{\tau}=\boldsymbol{\lambda}'\Lambda M^{\tau}\boldsymbol{m}_{1},
\end{aligned}
\tag{4.12}
$$

where the matrix $\Lambda=diag\left(m_{1,1},m_{1,2},m_{1,3},m_{1,4}\right)$. The autocorrelation function of price changes is given by:

$$
\zeta\left(\tau\right)=\frac{\boldsymbol{\lambda}'\Lambda M^{\tau}\boldsymbol{m}_{1}-\left(\boldsymbol{m}_{1}'\boldsymbol{\lambda}\right)^{2}}{\boldsymbol{m}_{2}'\boldsymbol{\lambda}-\left(\boldsymbol{m}_{1}'\boldsymbol{\lambda}\right)^{2}},
\tag{4.13}
$$

in our specific case $\zeta\left(\tau\right)=0$, because symmetry leads to $\boldsymbol{m}_{1}=0$.

Instead, the autocorrelation function of squared price changes $\rho(\tau)$ is:

$$\rho(\tau) = \frac{\boldsymbol{\lambda}' \Sigma M^{\tau} \boldsymbol{m}_2 - (\boldsymbol{m}_2' \boldsymbol{\lambda})^2}{\boldsymbol{m}_4' \boldsymbol{\lambda} - (\boldsymbol{m}_2' \boldsymbol{\lambda})^2}, \tag{4.14}$$

where the matrix $\Sigma = diag(m_{2,1}, m_{2,2}, m_{2,3}, m_{2,4})$. It is direct to show that $\rho(\tau)$ is an exponential function, $\exp(-a\tau)$, with $a = -\ln(p_{11} - p_{21})$,

As expected, both correlation functions depends on powers of the transition probability matrix M. For a Markov process, M is diagonalizable and we can write $M^{\tau} = C M_D^{\tau} C^{-1}$, where:

$$M_D^{\tau} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & (p_{11} - p_{21})^{\tau} \end{pmatrix},$$

$$C = \begin{pmatrix} 1 & 0 & 1 & 1 \\ \frac{p_{11}}{(p_{11}-1)} & 0 & 1 & \frac{p_{21}}{(p_{11}-1)} \\ 0 & 1 & 1 & 1 \\ 0 & \frac{p_{21}}{(p_{21}-1)} & 1 & \frac{p_{21}}{(p_{11}-1)} \end{pmatrix}.$$

In the limit case in which the spread is described by a Bernoulli process, the matrix M_B is not diagonalizable but has all eigenvalues in $\mathbb{R}$, i.e. $sp(M_B) = (0, 0, 0, 1)$, and we can compute its Jordan canonical form J_B. Thus we can rewrite the lag dependence as $M_B^{\tau} = E J_B^{\tau} E^{-1}$, where:

$$J_B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \; E = \begin{pmatrix} (p - p^2) & (1 - p^2) & p^2 & 0 \\ -p^2 & -p^2 & p^2 & 0 \\ (p - p^2) & -p^2 & p^2 & \frac{p-1}{p} \\ -p^2 & -p^2 & p^2 & 1 \end{pmatrix}.$$

The structure of block diagonal matrix J_B implies that $J_B^{\tau} = J_B^2 = 0, \; \forall \tau \geq 2$ and that $\rho(\tau)$ is a constant function for $\tau \geq 2$.

Discussion. The qualitative comparison of real data and model shows that the MS model is able to reproduce the distribution of price changes quite well. This can be seen by comparing Figs.4.1, 4.2 with Figs.4.5, 4.6. It is worth noting that, at least qualitatively, also the Bernoulli model MS_B is able to reproduce the underestimation of odd values of price changes with respect to the even values, as observed in real data. Therefore it is the coupling of spread and price change described by eq. 4.9, rather than the memory properties of spread, which is responsible of the behavior of the aggregated price change distribution of Fig. 4.2. The fact that price changes are uncorrelated is a consequence of our choice of symmetric conditional price changes, i.e. $\boldsymbol{m}_1 = 0$. Note however that the symmetry of unconditional price changes is not enough to guarantee lack of linear correlations. It is in fact possible to generalize the model of Eq. 4.9 in such a way that conditional price changes are not symmetric,

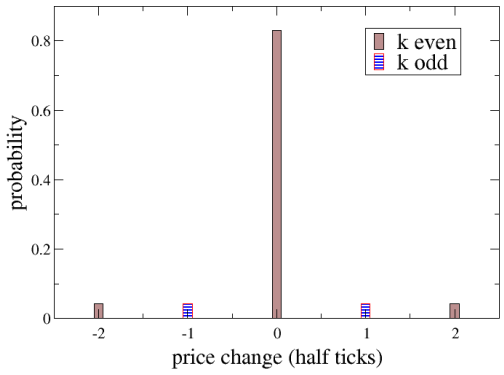

Figure 4.5: (Color online) Unconditional distribution of mid-price changes for the simulation of MS model calibrated on MSFT. The panel shows the histogram of the transaction to transaction change: $\Delta p\left(t, \Delta t = 1\right) = p_m\left(t + 1\right) - p_m\left(t\right)$.

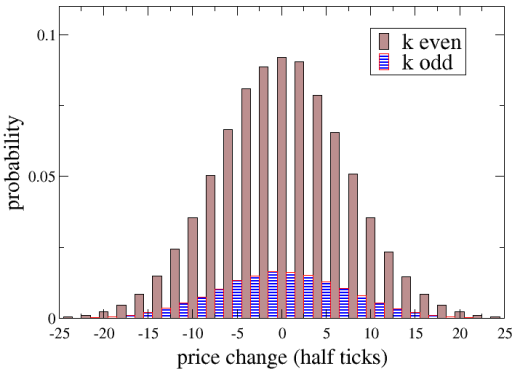

Figure 4.6: (Color online) Unconditional distribution of mid-price changes for the simulation of MS model calibrated on MSFT. The panel shows the histogram of the transaction to transaction change aggregated at 128 transactions: $\Delta p\left(t, \Delta t = 128\right) = p_m\left(t + 128\right) - p_m\left(t\right)$.

but unconditional price changes are symmetric, with exponentially decaying autocorrelation function of price changes. We investigate this case in section 4.9.3, where we study the autocorrelations between price changes. However the model fails to describe the volatility clustering. In fact, we have seen that the model has an exponentially decaying $\rho(\tau)$. Moreover, as the data calibration shows (see Fig. 4.7), the predicted behavior of $\rho(\tau)$ under the MS model is much smaller than the one observed in real data. Therefore this model is unable to reproduce the volatility clustering as well as any long memory property. This observation motivates us to develop a model that, preserving the structure of the coupling between spread and price changes discussed so far, is able to describe non exponential volatility clustering. This model is developed in the next section. 4.9.3

4.3.2 A double chain Markov model with logit regression

The Markov switching model is not able to explain the empirically observed correlation of squared price changes shown in Fig. 4.3. Therefore in the second class of models we consider an autoregressive switching model for price changes [99, 103] in order to study the correlation of squared price changes. The idea is to use logit regressions on past values of price changes and squared price changes. The model is thus defined by the following conditional probabilities:

$$\begin{aligned}
&P\left(\Delta p\left(t\right)|x\left(t\right)=k,\boldsymbol{\Omega}\left(t-1\right);\boldsymbol{\theta}_{k}\right),\ k\in\left\{1,2,3,4\right\};\\
&\boldsymbol{\Omega}'\left(t-1\right)=\Big(\Delta p^{2}\left(t-1\right),...,\Delta p^{2}\left(t-p\right),\\
&\Delta p\left(t-1\right),...,\Delta p\left(t-e\right)\Big)=\left(\boldsymbol{\Omega}'_{\Delta p^{2}},\boldsymbol{\Omega}'_{\Delta p}\right);\\
&\boldsymbol{\theta}'_{k}=\left(\alpha_{k},\boldsymbol{\beta}'_{k},\boldsymbol{\gamma}'_{k}\right),
\end{aligned}\tag{4.15}$$

where we define an informative $(p+e)$-dimensional vector of regressors $\boldsymbol{\Omega}$, made of the past e price changes and p squared price changes. Each parameter vector $\boldsymbol{\theta}_k$ is composed by the scalar α_k, the p-dimensional vector $\boldsymbol{\beta}_k$ which describes the regression on past values of squared price changes, and the e-dimensional vector $\boldsymbol{\gamma}_k$ which describes the regression on past price changes.

In order to handle the discreteness of price changes we make use of a logit regression.

To this end, we first convert the price changes series in a binary series $b(t) \in \{0,1\}$. When the spread remains constant between t and $t+1$ (i.e. $x(t)=1$ or $x(t)=4$), we set:

$$\begin{aligned}
\Delta p\left(t\right)=\pm2 \quad &\longrightarrow \quad b\left(t\right)=1,\\
\Delta p\left(t\right)=0 \quad &\longrightarrow \quad b\left(t\right)=0,
\end{aligned}\tag{4.16}$$

while when the spread changes, (i.e. $x(t) = 2$ or $x(t) = 3$) we set:

$$\begin{aligned} \Delta p\,(t) = 1 \quad &\longrightarrow \quad b\,(t) = 1, \\ \Delta p\,(t) = -1 \quad &\longrightarrow \quad b\,(t) = 0. \end{aligned} \tag{4.17}$$

Then, denoting by $\eta_k\,(t)$ the probability of having $b\,(t) = 1$ conditional on $x(t) = k$, the logit regression postulates that

$$\eta_k\,(t) = \frac{\exp\left(\alpha_k + \mathbf{\Omega}'_{\Delta p^2}\,(t-1)\,\boldsymbol{\beta}_k + \mathbf{\Omega}'_{\Delta p}\,(t-1)\,\boldsymbol{\gamma}_k\right)}{1 + \exp\left(\alpha_k + \mathbf{\Omega}'_{\Delta p^2}\,(t-1)\,\boldsymbol{\beta}_k + \mathbf{\Omega}'_{\Delta p}\,(t-1)\,\boldsymbol{\gamma}_k\right)}. \tag{4.18}$$

Therefore we obtain for the process[3] $\Delta p\,(t)$

$$\begin{aligned} &P\left(\Delta p\,(t) = \pm 2 | x\,(t) = 1, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_1\right) = \eta_1\,(t)\,/2, \\ &P\left(\Delta p\,(t) = 0 | x\,(t) = 1, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_1\right) = 1 - \eta_1\,(t)\,; \\ &P\left(\Delta p\,(t) = 1 | x\,(t) = 2, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_2\right) = \eta_2\,(t)\,, \\ &P\left(\Delta p\,(t) = -1 | x\,(t) = 2, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_2\right) = 1 - \eta_2\,(t)\,; \\ &P\left(\Delta p\,(t) = 1 | x\,(t) = 3, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_3\right) = \eta_3\,(t)\,, \\ &P\left(\Delta p\,(t) = -1 | x\,(t) = 3, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_3\right) = 1 - \eta_3\,(t)\,; \\ &P\left(\Delta p\,(t) = \pm 2 | x\,(t) = 4, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_4\right) = \eta_4\,(t)\,/2, \\ &P\left(\Delta p\,(t) = 0 | x\,(t) = 4, \mathbf{\Omega}\,(t-1)\,; \boldsymbol{\theta}_4\right) = 1 - \eta_4\,(t)\,. \end{aligned} \tag{4.19}$$

The intuition of the first two equations is the following. When between t and $t+1$ the spread remains equal to one tick ($x(t) = 1$, see left panel of Fig. 4), the probability that the midprice moves of two half ticks or remains the same depends on $\eta_1(t)$. This is a logistic function of a linear combination of the past e price changes and, more important, of past p squared price changes. Thus high volatile periods generate high volatility and for this reason the model is able to generate clustered volatility. Similar considerations holds for the other equations. Thus our model translate in a discrete logistic framework the self-excitation mechanism of volatility.

These equations define the general DCMM(e, p) model, because we can define a Markov process for $\Delta p\,(t)$. In the rest of the paper we will consider the case $e = 0$ and for the sake of simplicity we will denote DCMM(p)=DCMM$(0, p)$. In our case the independent latent Markov process is represented by the transition process $x\,(t)$ and the dependent Markov process is represented by the $\Delta p\,(t)$ processes. For the sake of clarity, here we consider the case $p = 1$, while its extension to a general value for p is considered in Appendix 4.8.

We define conditional Markov processes for $\Delta p\,(t)$ and $\Delta p^2\,(t)$ by means of the logit probabilities $\eta_k\,(t)$. The definition of the process for $\Delta p\,(t) \in \{-2, -1, 0, 1, 2\}$,

[3]Note that for the case $x(t) = 1, 4$ we have assumed equal probability for $\Delta p(t) = \pm 2$.

and $i, j \in \{1, 2, 3, 4, 5\}$, in the case of $p = 1$ (DCMM(1)) is the following:

$$P\left(\Delta p\left(t\right) = 3 - j | x\left(t\right) = k, \Delta p\left(t-1\right) = 3 - i;\ \boldsymbol{\theta}_k\right) = A_{k,ij}\ .$$

We have four possible transition matrices $A_{x(t)=k}$ for $k \in \{1, 2, 3, 4\}$, determined by the latent process $x\left(t\right)$. Their form is given in appendix 4.7.

Assuming that the latent process has reached the stationary distribution defined by Eq. 4.7, we can define an overall Markov chain by the transition matrix N that describes the $\Delta p\left(t\right)$ process:

$$N = \sum_{k=1}^{4} \lambda_k A_k. \tag{4.20}$$

The matrix N is defined by $6 + 4p$ parameters: $p_{11}, p_{21}, \alpha_k, \boldsymbol{\beta}'_k$.

Analogously the probabilities for the process $\Delta p^2\left(t\right) \in \{0, 1, 4\}$, and $i, j \in \{1, 2, 3\}$, in the case of $p = 1$ (DCMM(1)) is

$$P\Big(\Delta p^2\left(t\right) = \left(3 - j\right)^2 | x\left(t\right) = k, \Delta p^2\left(t-1\right) = \left(3 - i\right)^2; \boldsymbol{\theta}_k\Big) = V_{k,ij}\ .$$

which can be calculated from the matrix A (see appendix 4.7).

We can again define an overall Markov process for $\Delta p^2\left(t\right)$ described by a transition matrix S, assuming that the transition process $x\left(t\right)$ has reached the stationary distribution:

$$S = \sum_{k=1}^{4} \lambda_k V_k. \tag{4.21}$$

The matrix S is defined by $4 + 2p$ parameters: $p_{11}, p_{21}, \alpha_k, \boldsymbol{\beta}'_k$, where $k \in \{1, 4\}$, because $V_{k=2,3}$ do not depend by the past values of $\Delta p^2\left(t\right)$. In fact, the squared price change is a constant when $x\left(t\right) = 2, 3$, i.e. from eq. 4.19 we have $\Delta p^2\left(t\right) = 1\,, \forall t$.

The function $corr\left(\Delta p^2\left(t\right), \Delta p^2\left(t+\tau\right)\right) = \rho\left(\tau\right)$ for the DCMM(1) process is the correlation of the Markov(1) process defined by S. We solve the eigenvalue equation for S relative to the eigenvalue 1 in order to determine the stationary probability vector $\boldsymbol{\psi}$:

$$S'\boldsymbol{\psi} = \boldsymbol{\psi}. \tag{4.22}$$

If we define the vectors $\boldsymbol{\delta}$, $\boldsymbol{\delta}_2$ and $\boldsymbol{\xi}$, where $\delta_i = \left(3 - i\right)^2$, $\delta_{2,i} = \left(3 - i\right)^4$ and $\boldsymbol{\xi} = \boldsymbol{\delta} \odot \boldsymbol{\psi}$, the moments are given by:

$$\begin{aligned} E\left[\Delta p^2\left(t\right)\right] &= \boldsymbol{\delta}'\boldsymbol{\psi}, \\ E\left[\Delta p^4\left(t\right)\right] &= \boldsymbol{\delta}'_2\boldsymbol{\psi}, \\ E\left[\Delta p^2\left(t\right)\Delta p^2\left(t+\tau\right)\right] &= \boldsymbol{\xi}' S^\tau \boldsymbol{\delta}. \end{aligned} \tag{4.23}$$

Finally, we have the expression for $\rho\left(\tau\right)$ in the case $p = 1$:

$$\rho\left(\tau\right) = \frac{\boldsymbol{\xi}' S^\tau \boldsymbol{\delta} - \left(\boldsymbol{\delta}'\boldsymbol{\psi}\right)^2}{\boldsymbol{\delta}'_2\boldsymbol{\psi} - \left(\boldsymbol{\delta}'\boldsymbol{\psi}\right)^2}. \tag{4.24}$$

The generalization of the calculation of $\rho(\tau)$ to any value of the order p is reported in the appendix 4.8.

Estimation. In order to estimate the parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \boldsymbol{\theta}'_3, \boldsymbol{\theta}'_4)$ we maximize the following partial-loglikelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=p+1}^{T} \log\Bigg[\sum_{k=1}^{4} P(x(t) = k|\boldsymbol{\Omega}(t-1); \boldsymbol{\theta}_k) \\ P(b(t)|x(t) = k, \boldsymbol{\Omega}(t-1); \boldsymbol{\theta}_k)\Bigg], \tag{4.25}$$

where T is the length of sample, and we assume that parameters p_{11} and p_{21} are known. Since the dynamics of spread transitions is independent from the past informative set, i.e. $P(x(t) = k|\boldsymbol{\Omega}(t-1); \boldsymbol{\theta}_k) = P(x(t) = k)$, we have:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=p+1}^{T} \log\Bigg[\sum_{k=1}^{4} P(x(t) = k) \\ P(b(t)|x(t) = k, \boldsymbol{\Omega}(t-1); \boldsymbol{\theta}_k)\Bigg]. \tag{4.26}$$

In the case of large tick assets, it is $\lambda_1 \approx 1$ and we can use the approximation

$$\mathcal{L}(\boldsymbol{\theta}) \approx \sum_{t=p+1}^{T} \log\Big(P(b(t)|x(t) = 1, \boldsymbol{\Omega}(t-1); \boldsymbol{\theta}_1)\Big). \tag{4.27}$$

For example for MSFT we have $\lambda_1 \approx 0.9$. With this approximation we estimate only the vector $\boldsymbol{\theta}_1$ and the parameter θ_4 of Eq.4.9, that are enough in order to define matrices V_k. Moreover, we have the following approximation:

$$V_{x(t)=4} \approx \begin{pmatrix} 2\theta_4 & 0 & 1-2\theta_4 \\ 2\theta_4 & 0 & 1-2\theta_4 \\ 2\theta_4 & 0 & 1-2\theta_4 \end{pmatrix}.$$

In this way we neglect the contribution of regressors $\boldsymbol{\Omega}(t-1)$ (weighted by $\boldsymbol{\beta}_4$) and make use of the simpler expression in Eq. 4.9 when $x(t) = 4$. As before, this approximation holds if the weight of $V_{x(t)=4}$ is negligible, i.e. $\lambda_4 \approx 0$, i.e. when there is a small number of spread transitions $s(t) = 2 \to s(t+1) = 2$. This is the case when we have large tick assets, where we have almost always $s(t) = 1$. In the case of MSFT asset for example we have $\lambda_4 \simeq 0.04$.

We have performed the calculation of the autocorrelation $\rho(\tau)$ of the squared price changes for $p = 1, 3$ and the result is reported in Fig. 4.7. We have calibrated the parameters on the MSFT asset (see next Sections for details). We note that the MS and MS_B models underestimate very strongly $\rho(\tau)$. Note that for the MS model, $\rho(\tau)$ calibrated on real data is very small but not zero as predicted by the theory. The

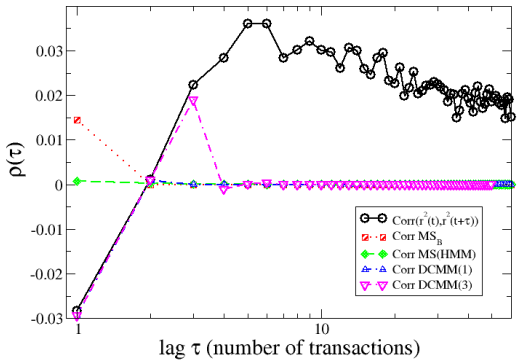

Figure 4.7: (Color online) Autocorrelation function of squared price changes $\rho(\tau)$, reported in $\log_{10}$-linear scales. The black circles are the real data of MSFT asset. The red squares are the result of the MS_B model, the green diamonds refer to the MS model, the blu up triangles refer to the DCMM(1) model and the pink down triangles refer to DCMM(3) model, all calibrated on the MSFT asset.

DCMM(p) model, on the other hand, is able to fit very well $\rho(\tau)$ up to lag $\tau = p$. Remarkably the model captures very well also the negative correlation for very short lags. However this observation indicates that an higher order DCMM(p) model might be able to fit better the real data. In the next Sections we will show that this is indeed the case.

4.4 Data

We have investigated two stocks, namely Microsoft (MSFT) and Cisco (CSCO), both traded at NASDAQ market in the period July-August 2009, corresponding to 42 trading days. Data contain time stamps corresponding to order executions, prices, size of trading volume and direction of trading. The time resolution is one millisecond. In this article we report mostly the results for MSFT asset, which are very similar to those for CSCO.

Non stationarities can be very important when investigating intraday financial data. For this reason and in order to restrict our empirical analysis to roughly stationary time intervals, we first compute the intensity of trading activity at time t conditional to a specific value k of mid-price change, i.e. $p(t|\Delta p(t) = k)$. As we can see from Figure 4.8, the unconditional trading intensity $p(t)$ is not stationary during the day [19]. Trading activity is very high at the beginning and at the end of the day and, as usually done to remove the effect of opening and closing auction, we discard transaction data in the first and last six minutes of trading day. Moreover the figure shows that the relative frequencies of the three values of price changes change during the day, except for price changes larger than two ticks that are very rare throughout the day. Most important, in the first part of the day, one tick or

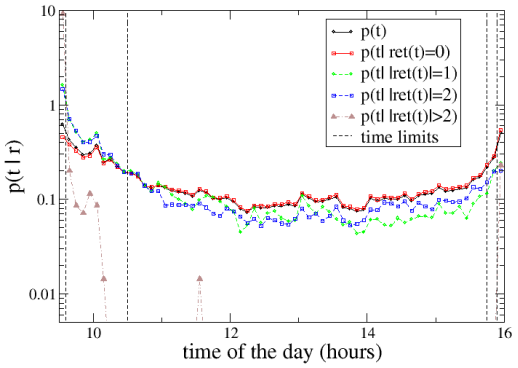

Figure 4.8: (Color online) Unconditional and conditional probability distributions describing the number of transaction events per unit time, reported in linear-log$_{10}$ scales. We bin recorded time events for MSFT into 6 minutes intervals.

asset	activity	# trades	mean (ticks/2)	σ	ex. ku.	$\hat{\pi}_1$
MSFT	high	184,542	$-2.82 * 10^{-4}$	0.652	5.13	0.92
42 days	low	348,253	$8.96 * 10^{-4}$	0.514	9.89	0.95
CSCO	high	145,084	$-1.32 * 10^{-3}$	0.673	4.73	0.92
42 days	low	275,879	$1.44 * 10^{-3}$	0.551	8.46	0.95

Table 4.1: Summary statistics for assets MSFT and CSCO for the two subsamples of high and low trading activity. σ is the standard deviation, ex. ku. is the excess kurtosis of transaction to transaction price changes and $\hat{\pi}_1$ is the fraction of time the spread is equal to one tick. The values of the mean are not significantly different from zero. A one sample t-test for the mean of price changes, performed on 42 trading days, does not reject the null-hypothesis of a zero value for the mean.

two tick price changes are more frequent than zero price changes, while after approximately $10:30$ the opposite is true. For this reason we split our times series in two subsamples. The first sample, corresponding to a period of high trading intensity, covers the time sets $t \in (9:36, 10:30) \cup (15:45, 15:54)$, where time is measured in hours. The second sample, corresponding to low trading intensity, covers the time set $t \in [10:30, 15:45]$. Table 1 reports a summary statistics of the two subsamples.

We then analyze the empirical autocorrelation function of squared price changes $corr\left(\Delta p^2(t), \Delta p^2(t+\tau)\right) = \rho(\tau)$ for these two series. As we can see from Fig. 4.9, for $\tau > 5$ transactions both time series display a significant positive and slowly decaying autocorrelation, which is a quantitative manifestation of volatility clustering. The series corresponding to low trading activity displays smaller, yet very persistent, volatility clustering.

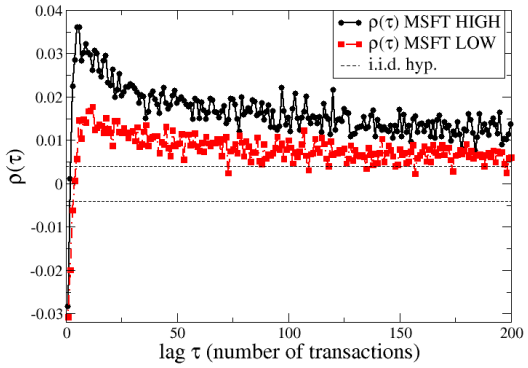

Figure 4.9: (Color online) Sample autocorrelation function of squared price changes, $\rho(\tau)$ for MSFT. Black circles refer to high trading activity series and the red squares refer to low trading activity series. The dashed lines indicate 2σ confidence intervals in the hypothesis of i.i.d. time series.

4.5 Estimation details and comparison with real data

We have estimated the models described in Secs.4.3.1 and 4.3.2 and we have used Monte Carlo simulations to generate artificial time series calibrated on real data. The statistical properties of these time series have been compared with those from real data.

More specifically we have considered three models: (i) the MS_B model, where spread is described by a Bernoulli process and there are no logit regressors; (ii) the MS model, where spread is a Markov(1) process and there are no logit regressors; (iii) the DCMM(p) model, where spread is a Markov(1) process and the set of logit regressors includes only the past p values of squared price changes. Notice therefore that in this last model we set $e = 0$. Finally, we have estimated the model separately for high and low activity regime.

4.5.1 Estimation of the models

From spread and price changes data we computed the MLE estimators $\hat{\pi}_1$, $\hat{p}_{11}$, $\hat{p}_{21}$, $\hat{\theta}_1$, $\hat{\theta}_4$ of the parameters defined in Sec.4.3.1. They are given by

$$\begin{aligned} \hat{\pi}_1 &= \frac{n_1}{N_s}, \\ \hat{p}_{ij} &= \frac{n_{ij}}{\sum_{j=1}^{2} n_{ij}}, \\ \hat{\theta}_k &= \frac{1}{2}\left(1 - \frac{n_{0k}}{N_k}\right), \end{aligned} \tag{4.28}$$

par.	MSFT high	MSFT low	CSCO high	CSCO low
$\hat{\pi}_1$	9.17×10^{-1}	9.52×10^{-1}	9.22×10^{-1}	9.54×10^{-1}
$\hat{p}_{11}$	9.53×10^{-1}	9.72×10^{-1}	9.53×10^{-1}	9.72×10^{-1}
$\hat{p}_{21}$	5.22×10^{-1}	5.50×10^{-1}	5.61×10^{-1}	5.84×10^{-1}
$\hat{\theta}_1$	4.81×10^{-2}	2.85×10^{-2}	5.15×10^{-2}	3.37×10^{-2}
$\hat{\theta}_4$	1.51×10^{-3}	2.65×10^{-4}	1.01×10^{-3}	0.00×10^{-4}

Table 4.2: Estimated parameters for the two subsamples of high and low trading activity for the MSFT and CSCO stocks.

where n_1 is the number of times $s(t) = 1$, N_s is the length of the spread time series, n_{ij} is the number of times the value of spread i is followed by the value j, n_{0k} is the number of times price changes are zero in the regime $x(t) = k$, and N_k is the length of the subseries of price changes in the same regime. For the last estimator $\hat{\theta}_k$ we count only zero price changes because we assumed that the price changes are distributed symmetrically in the set $(-2, 0, 2)$. We have checked that this assumption represents a good approximation for our data sets. The estimated parameters for MSFT asset are shown in Table 2.

In order to estimate the DCMM(p) model we need to estimate the vector $\boldsymbol{\theta}$. For both regimes we use the approximated log-likelihood of Eq. 4.27 because we have for low volatility series $P(x(t) = 1) \approx 0.92$ and for high volatility $P(x(t) = 1) \approx 0.87$. Thus we need to estimate only the vector $\boldsymbol{\theta}_1 = (\alpha_1, \boldsymbol{\beta_1}')$ by a standard generalized linear regression and we use an iterative reweighted least squares technique [146]. In this way we generate the price changes series in regime $x(t) = 1$, instead for the other regimes the generator follows the rules in Eq. 4.9, i.e. we use the estimator $\hat{\theta}_4$. The order of model is fixed to $p = 50$ in order to investigate the impact of past squared price changes on the price changes process. For simplicity we report here only the results from high activity time series.

We find $\alpha_1 = -2.921(0.019)$ and we report the first 25 values of β_{1i} in Table 3. The estimates of β_{1i} are significantly positive for $i > 2$ up to $i = 50$, with the exception of $i = 36, 37$. Moreover they display a maximum for $i = 6$. We perform a power law fit on these parameters, $\hat{\beta}_{1i} \propto i^{-\alpha}$, and we find a significant exponent $\alpha = 0.626(0.068)$. We hypothesize that this functional dependence of β_{1i} from i could be connected to the slow decay of the autocorrelation function of squared price changes, but we have not investigated further this aspect.

4.5.2 Comparison with real data

After having estimated the three models on the real data, we have generated for each model 25 data samples of length 10^6 observations. In this way we are able to determine an empirical statistical error on quantities that we measure on these artificial samples. We have considered three quantities to be compared with real data. Beside the autocorrelation of squared price changes, in order to analyze the price change distribution at different transaction time scales Δt, we have measured the empirical standard deviation and excess kurtosis:

i	$\hat{\beta}_{1i}$	st.error	$z-value$	$\hat{\beta}_{1i}$	st.error	$z-value$
	MSFT high	MSFT high	MSFT high	CSCO high	CSCO high	CSCO high
1	$-1.56*10^{-1}$	$9*10^{-3}$	-18.4 $***$	-2.85	$2.1*10^{-2}$	-136 $***$
2	$-4.03*10^{-2}$	$7.4*10^{-3}$	-5.45 $***$	$-1.69*10^{-1}$	$9.2*10^{-3}$	-18.5 $***$
3	$2.18*10^{-2}$	$7.0*10^{-3}$	3.12 $**$	$-2.88*10^{-2}$	$7.8*10^{-3}$	-3.72 $***$
4	$4.58*10^{-2}$	$6.9*10^{-3}$	6.61 $***$	$1.33*10^{-2}$	$7.5*10^{-3}$	1.77 .
5	$7.13*10^{-2}$	$6.8*10^{-3}$	10.5 $***$	$4.17*10^{-2}$	$7.4*10^{-3}$	5.63 $***$
6	$7.59*10^{-2}$	$6.8*10^{-3}$	11.2 $***$	$6.79*10^{-2}$	$7.3*10^{-3}$	9.36 $***$
7	$5.94*10^{-2}$	$6.9*10^{-3}$	8.57 $***$	$5.42*10^{-2}$	$7.4*10^{-3}$	7.33 $***$
8	$6.06*10^{-2}$	$6.9*10^{-3}$	8.76 $***$	$5.27*10^{-2}$	$7.4*10^{-3}$	7.12 $***$
9	$5.94*10^{-2}$	$6.9*10^{-3}$	8.55 $***$	$4.52*10^{-2}$	$7.5*10^{-3}$	6.05 $***$
10	$5.58*10^{-2}$	$7.0*10^{-3}$	8.01 $***$	$5.27*10^{-2}$	$7.4*10^{-3}$	7.11 $***$
11	$5.69*10^{-2}$	$6.9*10^{-3}$	8.20 $***$	$5.84*10^{-2}$	$7.4*10^{-3}$	7.91 $***$
12	$4.14*10^{-2}$	$7.1*10^{-3}$	5.86 $***$	$4.99*10^{-2}$	$7.5*10^{-3}$	6.71 $***$
13	$5.79*10^{-2}$	$6.9*10^{-3}$	8.36 $***$	$4.86*10^{-2}$	$7.5*10^{-3}$	6.52 $***$
14	$5.17*10^{-2}$	$7.0*10^{-3}$	7.40 $***$	$4.17*10^{-2}$	$7.5*10^{-3}$	5.55 $***$
15	$4.18*10^{-2}$	$7.1*10^{-3}$	5.93 $***$	$4.51*10^{-2}$	$7.5*10^{-3}$	6.03 $***$
16	$3.76*10^{-2}$	$7.1*10^{-3}$	5.30 $***$	$4.91*10^{-2}$	$7.4*10^{-3}$	6.61 $***$
17	$4.86*10^{-2}$	$7.0*10^{-3}$	6.92 $***$	$3.05*10^{-2}$	$7.6*10^{-3}$	4.02 $***$
18	$5.11*10^{-2}$	$7.0*10^{-3}$	7.31 $***$	$5.02*10^{-2}$	$7.4*10^{-3}$	6.75 $***$
19	$3.52*10^{-2}$	$7.1*10^{-3}$	4.95 $***$	$3.78*10^{-2}$	$7.5*10^{-3}$	5.01 $***$
20	$2.96*10^{-2}$	$7.2*10^{-3}$	4.14 $***$	$2.90*10^{-2}$	$7.6*10^{-3}$	3.81 $***$
21	$3.92*10^{-2}$	$7.1*10^{-3}$	5.54 $***$	$2.86*10^{-2}$	$7.6*10^{-3}$	3.77 $***$
22	$2.51*10^{-2}$	$7.2*10^{-3}$	3.49 $***$	$2.63*10^{-2}$	$7.6*10^{-3}$	3.46 $**$
23	$2.70*10^{-2}$	$7.2*10^{-3}$	3.76 $***$	$2.72*10^{-2}$	$7.6*10^{-3}$	3.58 $***$
24	$3.50*10^{-2}$	$7.1*10^{-3}$	4.93 $***$	$3.01*10^{-2}$	$7.6*10^{-3}$	3.98 $***$
25	$2.32*10^{-2}$	$7.2*10^{-3}$	3.23 $**$	$2.42*10^{-2}$	$7.6*10^{-3}$	3.17 $**$

Table 4.3: Estimated parameters β_{1i} for MSFT and CSCO asset in the high activity regime. Stars indicate significance levels: $***(0.001)$, $**(0.01)$, $*(0.05)$, $.(0.1)$, (1).

$$\sigma(\Delta t) = \left(E\left[((p_m(t+\Delta t) - p_m(t)) - E[p_m(t+\Delta t) - p_m(t)])^2\right]\right)^{1/2}, \tag{4.29}$$

$$\kappa(\Delta t) = \frac{E\left[((p_m(t+\Delta t) - p_m(t)) - E[p_m(t+\Delta t) - p_m(t)])^4\right]}{\sigma^4(\Delta t)} - 3\,. \tag{4.30}$$

The normalized standard deviation $\sigma_N(\Delta t) = \sigma(\Delta t)/\sqrt{\Delta t}$ gives information of the diffusive character of the price process, because $\sigma_N(\Delta t)$ is constant for diffusion. The behavior of $\kappa(\Delta t)$ as a function of Δt, instead, gives information about the presence or not of non-linear long range correlations between price changes [46]. We first investigate the autocorrelation properties of squared price changes $\rho(\tau)$. This function is compatible with zero for MS_B and MS models except for the first lag where we have measured a significant positive value, e.g. $\rho(\tau = 1) \approx 0.01$ in the case of MS_B for MSFT. The model with regressors DCMM(50), instead, is able to reproduce remarkably well the values of $\rho(\tau)$ up to $\tau = 50$, as we can see from Figs. 4.10, 4.11, both for MSFT and for CSCO. The behavior of $\rho(\tau)$ around $\tau \simeq 0$ is

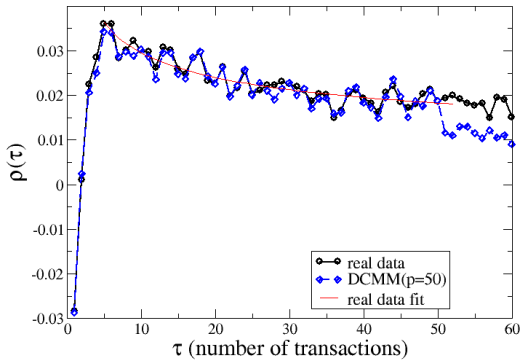

Figure 4.10: (Color online) Empirical autocorrelation functions $corr\left(\Delta p^2\left(t\right), \Delta p^2\left(t+\tau\right)\right)$ for real (black circles) and simulated (blue diamonds) data according to DCMM(50) model. The red line is a power law fit on the real data. The panel refers to the model fitted on MSFT data for high volatility series.

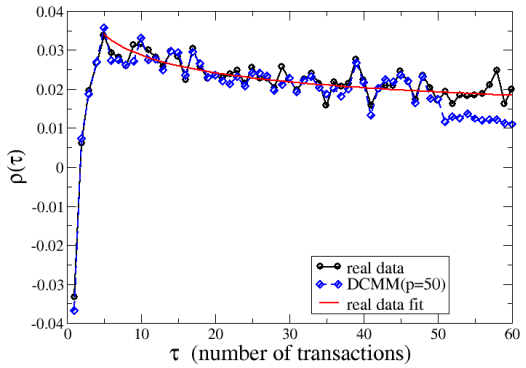

Figure 4.11: (Color online) Empirical autocorrelation functions $corr\left(\Delta p^2\left(t\right), \Delta p^2\left(t+\tau\right)\right)$ for real (black circles) and simulated (blue diamonds) data according to DCMM(50) model. The red line is a power law fit on the real data. The panel refers to the model fitted on CSCO data for high volatility series.

also very well reproduced by the model. The model underestimates the values of the autocorrelation of the real process for $\tau > 50$ but it generates values that are still significantly positive. We have performed a power law fit, i.e. $\rho\left(\tau\right) \propto \tau^{-\alpha}$, on real and DCMM(50) simulated data for values of lags corresponding to $\tau \in [6, 50]$. For real data we found $\alpha = 0.298(0.023)$ and for simulated data $\alpha = 0.300(0.028)$. Since $\alpha < 1$ this model is able to reproduce long memory shape of correlation $\rho\left(\tau\right)$ for a number of values of lags τ equal to the order of model p.

We then analyzed the distributional properties, i.e. normalized standard deviation $\sigma_N\left(\Delta t\right)$ and excess kurtosis $\kappa\left(\Delta t\right)$. For each value of Δt and for each model we

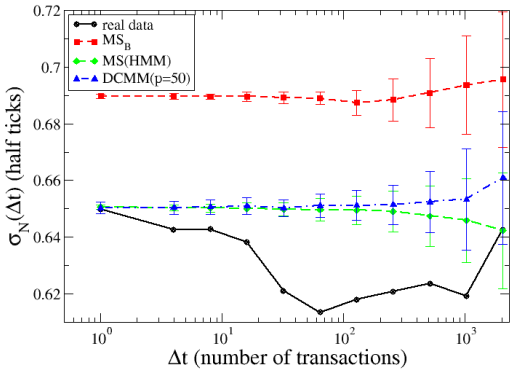

Figure 4.12: (Color online) Rescaled volatility, reported in $\log_{10}$-linear scales, $\sigma_N(\Delta t)$ of aggregated price changes on time scale Δt for MS_B (red line), MS (green line), and $DCMM(p = 50)$ (blue line), compared with the same quantity for MSFT data for high volatility series (black line). Error bars are the standard deviation obtained from 25 Monte Carlo simulations of the corresponding models.

calculate the average and standard deviation of the 25 simulations and we compare the simulation results with real data (see Figs. 4.12, 4.13).

The three models are clearly diffusive. This is a consequence of our modeling assumption, i.e. uncorrelatedness of price changes. The most important thing to observe is that MS and DCMM(50) models reproduce the empirical values of σ_N better than the MS_B model. The difference between MS and DCMM(50) models are appreciable only for $\Delta t > 128$, i.e. this parameter is almost the same for these two models. The behavior of excess kurtosis, instead, is different between the models (see Fig. 4.13). The excess kurtosis for MS_B and MS models is well fit by a power law $\kappa(\Delta t) \sim \Delta^{-\alpha}$ with $\alpha = 0.901(0.027)$ (MS_B) and $\alpha = 0.997(0.052)$ (MS). These values are consistent with a short range correlation of volatility. In fact, it can be shown [46] that stochastic volatility models with short range autocorrelated volatility are characterized by $\alpha = 1$. On the contrary, stochastic volatility models with long range autocorrelated volatility display a slower decay. This is exactly what it is observed for real data and for the DCMM(50) model. In both cases we observe an anomalous scaling of kurtosis that is more compatible with a stochastic volatility model in which volatility is a long memory process.

4.6 Conclusions

We have developed Markov switching models for describing the coupled dynamics of spread and price changes of large tick assets in transaction time. The underlying Markov process is the process of transitions between consecutive spread values. In this way price changes are described by different processes depending on whether the spread is constant or not in time. We have shown that this mechanism is needed

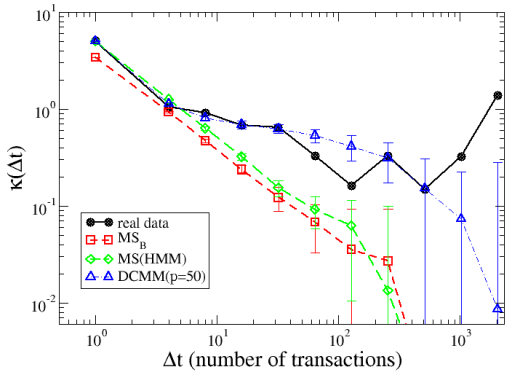

Figure 4.13: (Color online) Excess kurtosis, reported in $\log_{10}$-$\log_{10}$ scales, $\kappa(\Delta t)$ of aggregated price changes on time scale Δt for MS_B (red line), MS(green line), $DCMM(p = 50)$ (blue line), compared with the same quantity for MSFT data for high volatility series (black line). Error bars are the standard deviation obtained from 25 Monte Carlo simulations of the corresponding models.

in order to model the different shapes of the distribution of mid-price changes at different aggregation in number of trades, i.e. different time scales. In order to be able to model the persistent volatility clustering, we have introduced a Markov model with logit regressors represented by past values of price changes and squared price changes.

We have calibrated the model on NASDAQ stocks and we have found that the model reproduces remarkably well and in a quantitative way the empirical stylized facts. In particular we are able to reproduce the shape of the distribution at different time scales, uncorrelated price changes, diffusivity, slowly decaying autocorrelation function of squared price changes, and anomalous decay of kurtosis on different time scales.

As a possible extension, we observe that, if we want to reproduce more precisely the autocorrelation function of squared price changes up to a certain number of lags, we need to estimate a number of parameters, i.e. order of model, at least equal to this value. We find that these parameters scale with a power law function of parameter's index, i.e it is a function of the number of past lags at which regressors are defined. A possible improvement of this model could be to develop a model in which we estimate directly a parametric function with a small number of parameters (for example a power law function) that can describe how these parameters scale when we consider a certain order for the model.

A modeling approach based on an independent Markov process for the spread dynamics is only a simplified framework. Future research will be needed in order to investigate much sophisticated models that can allow long memory spread processes and a direct impact of price changes dynamics on the spread dynamics. This could be useful in order to extend our modeling approach to a broader class of assets like small tick assets.

4.7 Appendix 1: Matrices of the DCMM(1) model

There are four possible transition matrices $A_{x(t)=k}$ for $k \in \{1,2,3,4\}$, determined by the latent process $x(t)$:

$$A_{x(t)=1} = \begin{pmatrix} \eta_1\left(\Delta p^2(t-1)=4\right)/2 & 0 & 1-\eta_1\left(\Delta p^2(t-1)=4\right) & 0 & \eta_1\left(\Delta p^2(t-1)=4\right)/2 \\ \eta_1\left(\Delta p^2(t-1)=1\right)/2 & 0 & 1-\eta_1\left(\Delta p^2(t-1)=1\right) & 0 & \eta_1\left(\Delta p^2(t-1)=1\right)/2 \\ \eta_1\left(\Delta p^2(t-1)=0\right)/2 & 0 & 1-\eta_1\left(\Delta p^2(t-1)=0\right) & 0 & \eta_1\left(\Delta p^2(t-1)=0\right)/2 \\ \eta_1\left(\Delta p^2(t-1)=1\right)/2 & 0 & 1-\eta_1\left(\Delta p^2(t-1)=1\right) & 0 & \eta_1\left(\Delta p^2(t-1)=1\right)/2 \\ \eta_1\left(\Delta p^2(t-1)=4\right)/2 & 0 & 1-\eta_1\left(\Delta p^2(t-1)=4\right) & 0 & \eta_1\left(\Delta p^2(t-1)=4\right)/2 \end{pmatrix},$$

$$A_{x(t)=2} = \begin{pmatrix} 0 & \eta_2\left(\Delta p^2(t-1)=4\right) & 0 & 1-\eta_2\left(\Delta p^2(t-1)=4\right) & 0 \\ 0 & \eta_2\left(\Delta p^2(t-1)=1\right) & 0 & 1-\eta_2\left(\Delta p^2(t-1)=1\right) & 0 \\ 0 & \eta_2\left(\Delta p^2(t-1)=0\right) & 0 & 1-\eta_2\left(\Delta p^2(t-1)=0\right) & 0 \\ 0 & \eta_2\left(\Delta p^2(t-1)=1\right) & 0 & 1-\eta_2\left(\Delta p^2(t-1)=1\right) & 0 \\ 0 & \eta_2\left(\Delta p^2(t-1)=4\right) & 0 & 1-\eta_2\left(\Delta p^2(t-1)=4\right) & 0 \end{pmatrix},$$

where the temporal dependence is given only by the values of past squared price changes, i.e. $\Delta p^2(t-1)$, 1-step before the present time t. The others two matrices have the same definitions: $A_4 = A_1(\eta_1 \to \eta_4)$ and $A_3 = A_2(\eta_2 \to \eta_3)$. Here we are simply using Eq. 4.19 remembering that, when k is chosen, some states are forbidden.

In the case of $\Delta p^2(t)$, there are four possible transition matrices $V_{x(t)=k}$ for $k \in \{1,2,3,4\}$, determined by the latent process $x(t)$:

$$V_{x(t)=1} = \begin{pmatrix} \eta_1\left(\Delta p^2(t-1)=4\right) & 0 & 1-\eta_1\left(\Delta p^2(t-1)=4\right) \\ \eta_1\left(\Delta p^2(t-1)=1\right) & 0 & 1-\eta_1\left(\Delta p^2(t-1)=1\right) \\ \eta_1\left(\Delta p^2(t-1)=0\right) & 0 & 1-\eta_1\left(\Delta p^2(t-1)=0\right) \end{pmatrix}, \quad V_{x(t)=2} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The others two matrices have the same definitions: $V_4 = V_1(\eta_1 \to \eta_4)$ and $V_3 = V_2(\eta_2 \to \eta_3)$.

4.8 Appendix 2: Correlation of squared price changes for the DCMM(p) model

The definition of the Markov process for $\Delta p^2(t) \in \{0,1,4\}$ in the case of a general value of p for the DCMM model follows from Eq. 4.19. This stochastic process is a stationary Markov process of order p for each value of k [256] :

$$\begin{aligned} P\left(\Delta p^2(t) = (3-i_{p+1})^2 \,|x(t)=k;\, \Delta p^2(t-1) = (3-i_p)^2,\right. \\ \left. \cdots, \Delta p^2(t-p) = (3-i_1)^2; \boldsymbol{\theta}_k\right) = V_{x(t);i_1 i_2 \ldots i_{p+1}}, \end{aligned} \tag{4.31}$$

where we have $k \in \{1,2,3,4\}$ and a $p+1$-dimensional vector of indices $\hat{\boldsymbol{i}} = (i_1, i_2, \cdots, i_{p+1})$, where each index can assume values $i_l \in \{1,2,3\}$ for each $l \in \{1,2,\cdots,p+1\}$. We stress the concept that the index i_{p+1} defines the present value of the squared price change $\Delta p^2(t)$, instead the indices $i_1, \cdots, i_p$ define

the past history of the process of squared price changes, i.e. i_1 defines the oldest value of $\Delta p^2 = \Delta p^2 (t-p)$. The transition probabilities are given by:

$$\begin{aligned}
&V_{x(t)=k\in\{1,4\};i_1 i_2 ...,i_{p+1}=1} = \eta_k (i_1, \cdots, i_p) = \\
&= \frac{exp\left[\alpha_k + \sum_{l=1}^{p} \beta_{k,l} (3 - i_{p-l+1})^2\right]}{1 + exp\left[\alpha_k + \sum_{l=1}^{p} \beta_{k,l} (3 - i_{p-l+1})^2\right]}, \\
&V_{x(t)=k\in\{1,4\};i_1 i_2 ...,i_{p+1}=2} = 0, \\
&V_{x(t)=k\in\{1,4\};i_1 i_2 ...,i_{p+1}=3} = \\
&= \frac{1}{1 + exp\left[\alpha_k + \sum_{l=1}^{p} \beta_{k,l} (3 - i_{p-l+1})^2\right]}, \\
&V_{x(t)=k\in\{2,3\};i_1 i_2 ...,i_{p+1}=1} = 0, \\
&V_{x(t)=k\in\{2,3\};i_1 i_2 ...,i_{p+1}=2} = 1, \\
&V_{x(t)=k\in\{2,3\};i_1 i_2 ...,i_{p+1}=3} = 0,
\end{aligned} \tag{4.32}$$

for each value of the p-dimensional vector $\boldsymbol{i} = (i_1, \cdots, i_p)$. We have 3^{p+1} values for the transition probabilities with normalization:

$$\forall k;\ \forall\ i_1, \cdots, i_p\ :\ \sum_{i_{p+1}=1}^{3} V_{x(t)=k;i_1 i_2 ...,i_{p+1}} = 1. \tag{4.33}$$

We can recover an equivalent Markov(1) process defined on vector-states $\boldsymbol{Y}(t)$. We define a p-dimensional vector of squared price changes:

$$\begin{aligned}
\boldsymbol{Y}(t)[\boldsymbol{i}] \ &= \ \Big(\Delta p^2 (t-p+1) = (3-i_1)^2, \cdots \\
&\qquad \cdots, \Delta p^2 (t) = (3-i_p)^2\Big).
\end{aligned} \tag{4.34}$$

In this case the index i_p defines the present state of the squared price change $\Delta p^2 (t)$. The vector-process $\boldsymbol{Y}(t)$ is a first order Markov chain on the state space $\{0, 1, 4\}^p$, i.e. $\boldsymbol{Y}(t)$ can assume 3^p different values. We define four transition matrices $U_{x(t)=k} \in M_{3^p,3^p}(\mathbb{R})$ in order to represent the equivalent Markov process for each possible value of $x(t)$. These matrices describe the transition $\boldsymbol{Y}(t) \to \boldsymbol{Y}(t+1)$, that we could represent also by the transition between vectors of indices: $(i_1, \cdots, i_p) \to (i_2, \cdots, i_{p+1})$. We have to map the transition probabilities $V_{x(t)=k;i_1 i_2 ... i_{p+1}}$ to the elements of matrix

$U_{k;m,n}$, where $m,n \in \{1,\cdots,3^p\}$. We can obtain this by the following mappings:

$$\begin{aligned}
(i_1,\cdots,i_{p+1}) &\rightarrow (m,n)\,, \\
m\,(i_1,\cdots,i_p) &= \left[\sum_{l=1}^{p-1} 3^{p-l}\,(3-i_l)\right] + 4 - i_p, \\
n\,(i_2,\cdots,i_{p+1}) &= \left[\sum_{l=1}^{p-1} 3^{p-l}\,(3-i_{l+1})\right] + 4 - i_{p+1}, \\
U_{x(t)=k;m,n} &= V_{x(t)=k;i_1 i_2 \ldots i_{p+1}}.
\end{aligned} \tag{4.35}$$

This rules are unable to fill the entire matrix $U_{k;m,n}$, because when we study the Markov process for $\boldsymbol{Y}\,(t)$ we have a lot of forbidden transitions, so the elements of matrix that aren't captured by the above rules have 0 values. For the case $p=2$ the shape of U_k is:

$$U_1 = \begin{pmatrix}
[1-\eta_1\,(0,0)] & 0 & \eta_1\,(0,0) & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & [1-\eta_1\,(0,1)] & 0 & \eta_1\,(0,1) & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & [1-\eta_1\,(0,4)] & 0 & \eta_1\,(0,4) \\
[1-\eta_1\,(1,0)] & 0 & \eta_1\,(1,0) & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & [1-\eta_1\,(1,1)] & 0 & \eta_1\,(1,1) & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & [1-\eta_1\,(1,4)] & 0 & \eta_1\,(1,4) \\
[1-\eta_1\,(4,0)] & 0 & \eta_1\,(4,0) & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & [1-\eta_1\,(4,1)] & 0 & \eta_1\,(4,1) & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & [1-\eta_1\,(4,4)] & 0 & \eta_1\,(4,4)
\end{pmatrix},$$

$$U_2 = \begin{pmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{pmatrix}, \quad U_3 = U_2, \quad U_4 = U_1\,(\eta_1 \rightarrow \eta_4)\,.$$

In U_1 we have $\eta_1\,(i_1,i_2) = \eta_1\left(\Delta p^2\,(t-2) = (3-i_1)^2\,,\ \Delta p^2\,(t-1) = (3-i_2)^2\right)$. Finally, we define an overall Markov process for $\boldsymbol{Y}\,(t)$, defined by $4+2p$ parameters: $p_{11}, p_{21}, \alpha_k, \boldsymbol{\beta}'_k$, where $k \in \{1,4\}$:

$$S = \sum_{k=1}^{4} \lambda_k U_k, \tag{4.36}$$

where λ_k are given by Eq. 4.7.

Now our goal is to calculate the moments for the variable $\Delta p^2(t)$ from the process defined by Eq. 4.36. First of all we have to solve the eigenvalue equation for S relative to the eigenvalue 1 in order to determine the stationary probability vector for $\boldsymbol{Y}(t)$:

$$S'\boldsymbol{\Psi} = \boldsymbol{\Psi}. \tag{4.37}$$

The 3^p-dimensional vector $\boldsymbol{\Psi}$ represents all possible values of the stationary distribution of the variable $\boldsymbol{Y}(t)$:

$$P\left(\boldsymbol{Y}(t)\left[i_1, \cdots, i_p\right]\right) = \Psi_{m(i_1,\cdots,i_p)}. \tag{4.38}$$

From the 3^p-dimensional vector $\boldsymbol{\Psi}$ we compute the stationary 3-dimensional probability vector $\boldsymbol{\psi}' = (\psi_1, \psi_2, \psi_3)$ for the process $\Delta p^2(t)$, i.e. we have for each index $i_p \in \{1, 2, 3\}$:

$$\psi_{i_p} = P\left[\Delta p^2(t) = (3 - i_p)^2\right] = \sum_{i_1=1}^{3} \cdots \sum_{i_{p-1}=1}^{3} \Psi_{m(i_1,\cdots,i_p)}, \tag{4.39}$$

where i_p defines the present value of $\Delta p^2(t)$ and we use mappings defined in Eq. 4.35. The stationary probability to have a fixed value of Δp^2 at time t depends on all possible values of Δp^2 during the past $p-1$ lags. In order to determine the present probabilities we have to sum probabilities corresponding to all possible past trajectories defined by the past $p-1$ lags.

We compute $corr\left(\Delta p^2(t), \Delta p^2(t+\tau)\right) = \rho(\tau)$ by means of the transition probabilities $P\left(\Delta p^2(t) = (3-a)^2,\ \Delta p^2(t+\tau) = (3-b)^2\right)$, where $a, b \in \{1, 2, 3\}$, of the p-order Markov process in term of the matrix S:

$$\begin{aligned} &P\left(\Delta p^2(t) = (3-a)^2, \Delta p^2(t+\tau) = (3-b)^2\right) = \\ &= P\left(\boldsymbol{i}(a), \boldsymbol{j}(b)\right), \end{aligned} \tag{4.40}$$

where $\boldsymbol{i}(a) = (i_1, \cdots, i_p = a)$ and $\boldsymbol{j}(b) = (j_1, \cdots, j_p = b)$ are the p-dimensional vectors of indices describing the past $p-1$ lags respect to times t and $t+\tau$. We have to perform the sum of probabilities corresponding to each of the possible values of $i_1, \cdots, i_{p-1}$ and $j_1, \cdots, j_{p-1}$, i.e. on $i_l, j_l \in \{1, 2, 3\}$, $\forall l \in \{1, \cdots, p-1\}$. We use mappings defined in Eq. 4.35 and the matrix power S^τ, because we sum on all possible transitions $\boldsymbol{Y}(t) \rightarrow \boldsymbol{Y}(t+\tau)$ holding fixed the values of indices $i_p = a$ and $j_p = b$:

$$\begin{aligned} P\left(\boldsymbol{i}(a), \boldsymbol{j}(b)\right) &= \sum_{(i_1,\cdots,i_{p-1},i_p=a)} \sum_{(j_1,\cdots,j_{p-1},j_p=b)} P\Big(\boldsymbol{Y}(t)\left[\boldsymbol{i}(a)\right], \boldsymbol{Y}(t+\tau)\left[\boldsymbol{j}(b)\right]\Big) \\ &= \sum_{(i_1,\cdots,i_{p-1},i_p=a)} \sum_{(j_1,\cdots,j_{p-1},j_p=b)} \left(S^\tau\right)_{m(\boldsymbol{i}(a)),n(\boldsymbol{j}(b))} \Psi_{m(\boldsymbol{i}(a))}. \end{aligned} \tag{4.41}$$

The moments of our interest are:

$$\begin{aligned} E\left[\Delta p^2(t)\right] &= \sum_{i=1}^{3}(3-i)^2\psi_i = 4\psi_1+\psi_2, \\ E\left[\Delta p^4(t)\right] &= \sum_{i=1}^{3}(3-i)^4\psi_i = 16\psi_1+\psi_2, \\ E\left[\Delta p^2(t)\Delta p^2(t+\tau)\right] &= \sum_{a=1}^{3}\sum_{b=1}^{3}\Big[(3-a)^2(3-b)^2 \\ &\quad P(\boldsymbol{i}(a),\boldsymbol{j}(b))\Big], \end{aligned} \tag{4.42}$$

from which we can determine the function $\rho(\tau)$. We have determined the function $\rho(\tau)$ for $p=3$ making the following approximation for V_4:

$$\begin{aligned} V_{x(t)=k=4;i_1i_2...i_{p+1}=1} &= 2\theta_4, \\ V_{x(t)=k=4;i_1i_2...i_{p+1}=2} &= 0, \\ V_{x(t)=k=4;i_1i_2...i_{p+1}=3} &= 1-2\theta_4, \end{aligned} \tag{4.43}$$

this approximation is justified only in the case $\lambda_1 \approx 1$, i.e. we have the same approximation that leads us to Eq. 4.27. In this way we have found the results reported in Fig. 4.7 for DCMM($p=3$).

4.9 Appendix 3: Discussion on some simplifying modeling assumptions.

4.9.1 Using a Markovian spread

The level of $\rho(\tau)$ reported in fig. 4.3 is not a consequence of a non Markovian spread dynamics. To show this we have performed Montecarlo simulations of the MS model where we use as latent process the empirically observed spread, instead of a Bernoulli or Markov process. The left panel of Fig. 4.14 shows the empirical autocorrelation function of the spread. It clearly shows a very slow decay, possibly consistent with long memory, see Section 3.3 for more details. When used as an input of our model, the right panel of fig. 4.14 shows that the simulated autocorrelation of squared price changes (red line) is much smaller than the one empirically observed (black line). Therefore the slowly decaying autocorrelation of spread is not the origin of the empirically observed $\rho(\tau)$. The MS model is unable to reproduce this emprical fact, and for this reason we have introduced the DCMM model (which is able to do it).

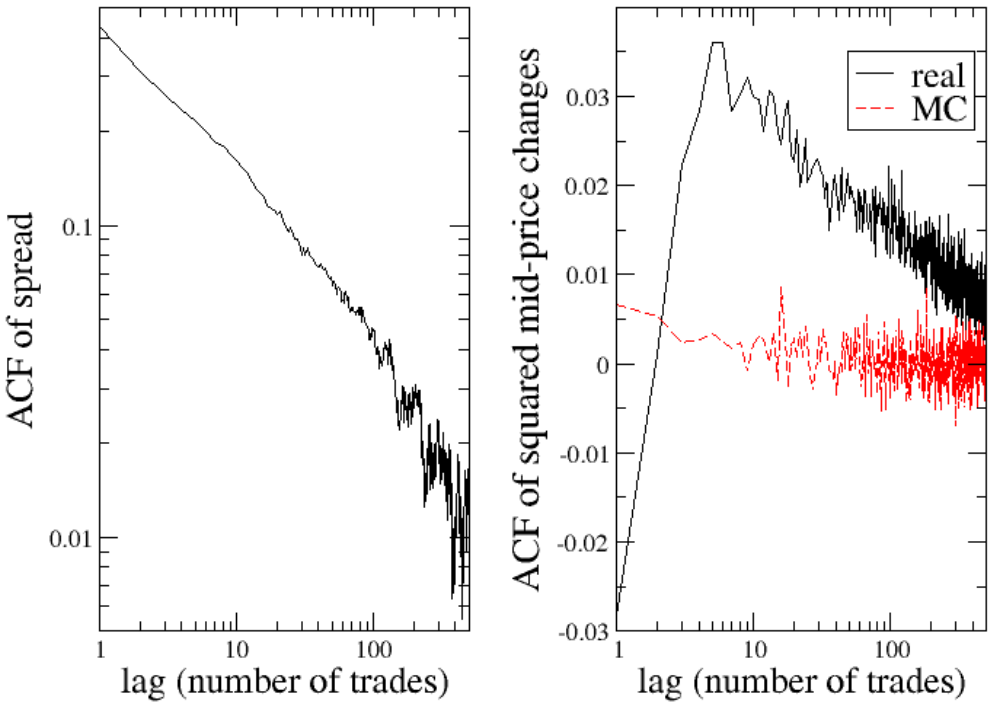

Figure 4.14: On the left panel we report the autocorrelation of the observed spread process. On the right we report the autocorrelation of the real series of squared price changes and that relative to our Montecarlo simulation, where the switching process is the real spread process. Both are computed for MSFT stock during the period of high activity. These results are the same for the other series of interest.

4.9.2 Conditioning spread on price changes

The model we propose is designed to describe the coupling between price changes and spread transitions. We did so by assuming the spread as an independent process and the price change as conditional to spread. We do not explicitly model conditional dependence of spread on price changes. However, the model is able to reproduce these conditional dependence as observed in the real data.

To show the conditional dependence of spread on price changes, in table 4.4 we report the first two empirical moments of the spread $s\left(t+1\right)$ conditional on the mid-price change $\Delta p = p_m\left(t+1\right) - p_m\left(t\right)$. When the price change is even, i.e. $\Delta p(t) = 0, 2, -2$, the mean spread is slightly above 1-tick, while when the price change is odd, i.e. $\Delta p(t) = 1, -1$, the mean spread is close to 1.5-tick.

In order to reproduce quantitatively the conditional mean and fluctuations of the spread, we have performed Montecarlo simulations of the models defined in the paper. We have also considered a model where the spread is given by the time series

moments	MSFT high	MSFT low	CSCO high	CSCO low
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=0\right]$	1.05	1.02	1.04	1.02
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=1\right]$	1.47	1.48	1.50	1.53
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=-1\right]$	1.53	1.51	1.51	1.47
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=2\right]$	1.00	1.00	1.00	1.00
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=-2\right]$	1.00	1.00	1.00	1.00
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=0)$	0.21	0.15	0.20	0.15
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=1)$	0.50	0.50	0.50	0.50
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=-1)$	0.50	0.50	0.50	0.50
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=2)$	0.04	0.00	0.04	0.00
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=-2)$	0.03	0.01	0.04	0.02

Table 4.4: We report the mean and the standard deviation of the spread distribution conditioned to price changes. The spread is measured in units of ticks. They are dependent from the value of the price change. The price changes, instead, are measured in units of half tick.

moments	MSFT high	MS_B	MS	DCMM(50)	real spread
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=0\right]$	1.05	1.01	1.05	1.05	1.05
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=1\right]$	1.47	1.50	1.50	1.50	1.51
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=-1\right]$	1.53	1.50	1.50	1.50	1.49
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=2\right]$	1.00	1.00	1.00	1.00	1.00
$\mathbb{E}\left[s\left(t+1\right) \mid \Delta p\left(t\right)=-2\right]$	1.00	1.00	1.00	1.00	1.00
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=0)$	0.21	0.09	0.21	0.21	0.21
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=1)$	0.50	0.50	0.50	0.50	0.50
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=-1)$	0.50	0.50	0.50	0.50	0.50
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=2)$	0.04	0.01	0.03	0.03	0.04
$\text{stdev}(s\left(t+1\right) \mid \Delta p\left(t\right)=-2)$	0.03	0.02	0.04	0.04	0.04

Table 4.5: We report the mean and the standard deviation of the empirical spread distribution conditioned to price changes for the MSFT high series against the values computed by Montecarlo simulations of four different models. The spread is measured in units of ticks. They are dependent from the value of the price change. The price changes, instead, are measured in units of half tick.

observed in the real data, i.e. a model where the switching process is a long memory process. In the tables this latter case is indicated with real spread. We illustrate the details for the MSFT stock. The results, reported in table 4.5 and in table 4.6, show that the Bernoullian model, MS_B , is unable to reproduce the conditional fluctuations of the spread, while the Markovian MS model is sufficient to reproduce the properties of the conditional distribution of spreads. The more complex models, DCMM and real data, perform like the Markovian model. In conclusion, even if we do not model explicitly the spread dynamics conditional to price changes, the conditional dependence of price changes on spread and the Markov memory of the spread process are enough to reproduce satisfactorily the observed spreads conditional to price changes. The additional contribution of an explicit conditional dependence of spread on price change is therefore small and for this reason we neglect it.

moments	MSFT low	MS_B	MS	DCMM(50)	real spread
$\mathbb{E}[s(t+1) \mid \Delta p(t)=0]$	1.02	1.00	1.02	1.02	1.02
$\mathbb{E}[s(t+1) \mid \Delta p(t)=1]$	1.48	1.50	1.50	1.50	1.50
$\mathbb{E}[s(t+1) \mid \Delta p(t)=-1]$	1.51	1.50	1.50	1.50	1.50
$\mathbb{E}[s(t+1) \mid \Delta p(t)=2]$	1.00	1.00	1.00	1.00	1.00
$\mathbb{E}[s(t+1) \mid \Delta p(t)=-2]$	1.00	1.00	1.00	1.00	1.00
$\mathrm{stdev}(s(t+1) \mid \Delta p(t)=0)$	0.15	0.05	0.15	0.15	0.15
$\mathrm{stdev}(s(t+1) \mid \Delta p(t)=1)$	0.50	0.50	0.50	0.50	0.50
$\mathrm{stdev}(s(t+1) \mid \Delta p(t)=-1)$	0.50	0.50	0.50	0.50	0.50
$\mathrm{stdev}(s(t+1) \mid \Delta p(t)=2)$	0.00	0.01	0.01	0.01	0.02
$\mathrm{stdev}(s(t+1) \mid \Delta p(t)=-2)$	0.01	0.01	0.01	0.01	0.02

Table 4.6: We report the mean and the standard deviation of the spread distribution conditioned to price changes for the MSFT low series against the values computed by Montecarlo simulations of four different models. The spread is measured in units of ticks. They are dependent from the value of the price change. The price changes, instead, are measured in units of half tick.

4.9.3 Autocorrelation of price changes

In this section we show some empirical facts regarding the autocorrelation of unconditional price changes and conditional price changes to the value of the transition $x(t)$. The empirical autocorrelation of price changes[4] is reported in fig. 4.15. The autocorrelation is not significant for almost all the time lags. The MSFT and CSCO high volatility series display a significant first lag autocorrelation. We have to note that in this case any autocorrelation between price changes is not due to the bid-ask bounce [134] because we are studying the midprice.

The Markov models defined by the eq. 4.9 do not allow any autocorrelation between price changes. These models can allow for autocorrelation between price changes if the conditional distribution of price changes are asymmetric with respect to the zero value. The following generalization of the model can describe exponentially autocorrelated price changes and asymmetric unconditional price changes distributions

$$\begin{aligned}
P(\Delta p(t)=\pm 2|x(t)=1;\boldsymbol{\theta}) &= \theta_1, \\
P(\Delta p(t)=0|x(t)=1;\boldsymbol{\theta}) &= 1-2\theta_1, \\
P(\Delta p(t)=1|x(t)=2;\epsilon_1) &= 1/2-\epsilon_1, \\
P(\Delta p(t)=-1|x(t)=2;\epsilon_1) &= 1/2+\epsilon_1, \\
P(\Delta p(t)=1|x(t)=3;\epsilon_2) &= 1/2+\epsilon_2, \\
P(\Delta p(t)=-1|x(t)=3;\epsilon_2) &= 1/2-\epsilon_2, \\
P(\Delta p(t)=\pm 2|x(t)=4;\boldsymbol{\theta}) &= \theta_4, \\
P(\Delta p(t)=0|x(t)=4;\boldsymbol{\theta}) &= 1-2\theta_4.
\end{aligned} \tag{4.44}$$

[4]We report the sample estimate according to the Box-Jenkins methodology [47] of the autocorrelation $\zeta(\tau)=\mathbb{E}\left[\left(\Delta p(t)-\mathbb{E}\left[\Delta p(t)\right]\right)\left(\Delta p(t+\tau)-\mathbb{E}\left[\Delta p(t)\right]\right)\right]/Var\left[\Delta p(t)\right]$.

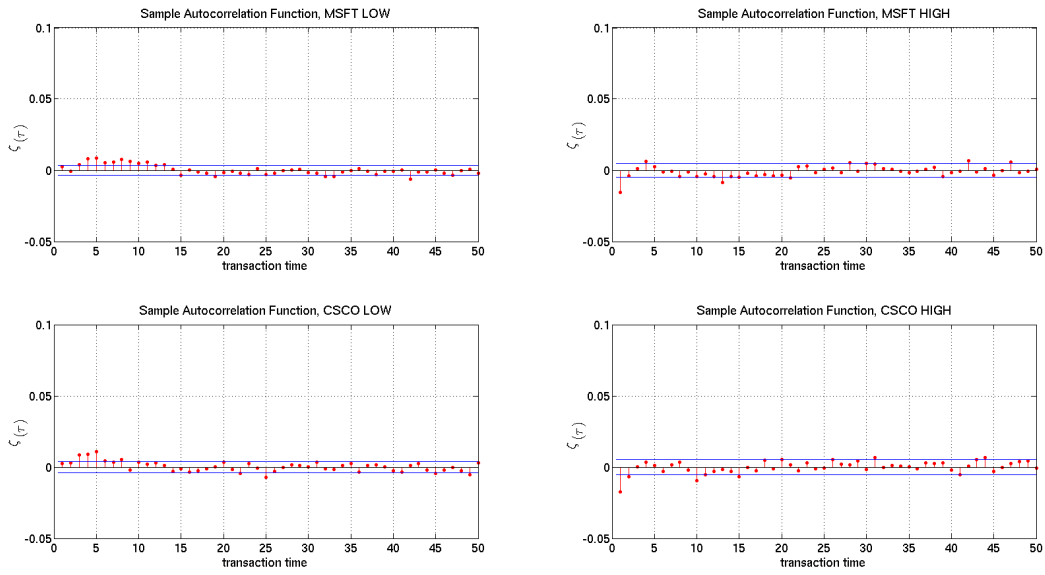

Figure 4.15: Empirical autocorrelation of price changes for the four time series under exam.

probability	MSFT low	MSFT high	CSCO low	CSCO high
$P(\Delta p\,(t) = 1 \vert x\,(t) = 2)$	0.48	0.48	0.47	0.48
$P(\Delta p\,(t) = -1 \vert x\,(t) = 2)$	0.52	0.52	0.53	0.52
$P(\Delta p\,(t) = 1 \vert x\,(t) = 3)$	0.51	0.53	0.52	0.51
$P(\Delta p\,(t) = -1 \vert x\,(t) = 3)$	0.49	0.47	0.48	0.49

Table 4.7: Conditional distribution of price changes to $x\,(t) = 2, 3$, i.e. when the spread changes during a trade.

where ϵ_1, ϵ_2 describe the perturbation with respect to the original model. The empirical asymmetry, i.e. represented by ϵ, is of the order of 1% as we can observe in tab. 4.7. When the spread does not change during the transition, the conditional distribution does not show appreciable asymmetries. We choose to use the same value for ϵ_1 and ϵ_2, i.e. the mean $(\epsilon_1 + \epsilon_2)/2$. In this way we allow for symmetric unconditional distribution and exponentially autocorrelated price changes. This model implies symmetric unconditional distribution for price changes, because the stationary probability vector $\boldsymbol{\lambda}$ of the underlying Markov process for transition has $\lambda_2 = \lambda_3$. This property of the model reproduces empirical observations, i.e the asymmetry observed in the empirical unconditional distributions is not statistically significant. To motivate this statement we have computed the skewness and the difference between mean and median of 42 different sub-samples of the original time series of price changes, where each sub-sample corresponds to one of the 42 trading days. A one-sample t-test, reported in tab. 4.8, shows that our data are consistent with the hypothesis of symmetric unconditional distributions. We show the autocorrelation implied by the model of eq. 4.44 in the case of MSFT high, i.e. $\epsilon_1 = \epsilon_2 = 0.025$. This choice is motivated

moments	MSFT high	MSFT low	CSCO high	CSCO low
mean	0.9513	0.4249	0.1857	0.1598
skewness	0.5628	0.3581	0.1133	0.2536
mean-median	0.9513	0.4249	0.1857	0.1598

Table 4.8: We report the p-value relative to the one sample t-test for the mean, skewness and mean-median computed on 42 trading days. The null-hypothesis is that the population mean is zero.

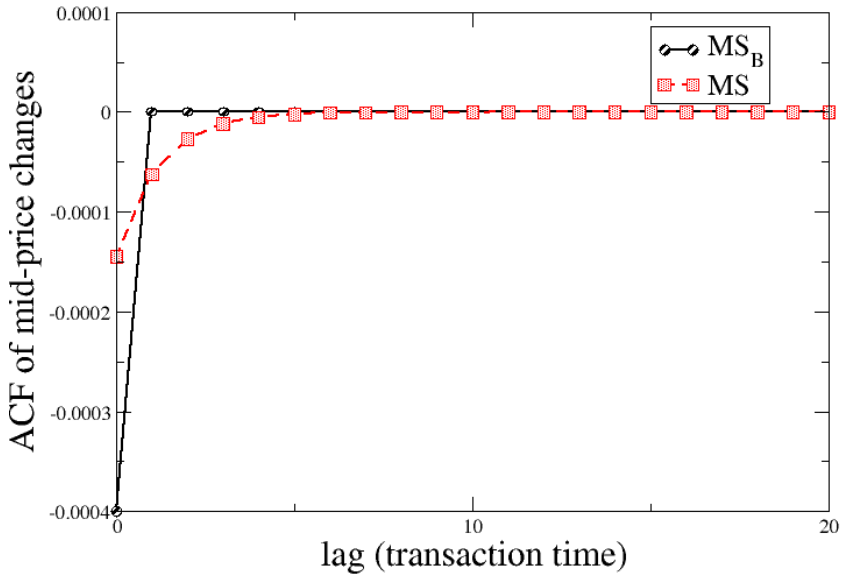

Figure 4.16: Autocorrelation of price changes for the MS and MS_B models for $\epsilon = 0.025$.

by the fact that only high volatility time series appear to show a significant first lag autocorrelation. The result given by eq. 4.13 is shown in fig. 4.16. The values of autocorrelation are within the error bars of fig. 4.15, i.e. the region compatible with uncorrelated returns. Therefore we can conclude that such models are unable to reproduce the first lag autocorrelation found in the case of high volatility series.

It is clear from such discussion that Markov switching models do not allow to describe the presence of a first lag autocorrelation for price changes. The following observations lead us to think that in order to reproduce the autocorrelation between price changes we should consider to model the autocorrelation conditioned to specific spread transitions. The behavior of correlations conditioned to specific values of transitions $x(t)$, i.e. $corr\left(\Delta p(t), \Delta p(t+\tau) | x(t) = i\right)$, $i \in \{1, 2, 3, 4\}$, is completely different from what found in the unconditional case. We report in fig. 4.17 the case of MSFT in the low regime, the other cases show a similar behavior. When the value of transitions is $x(t) = 2, 3$, i.e. when the spread is closing or widening, the

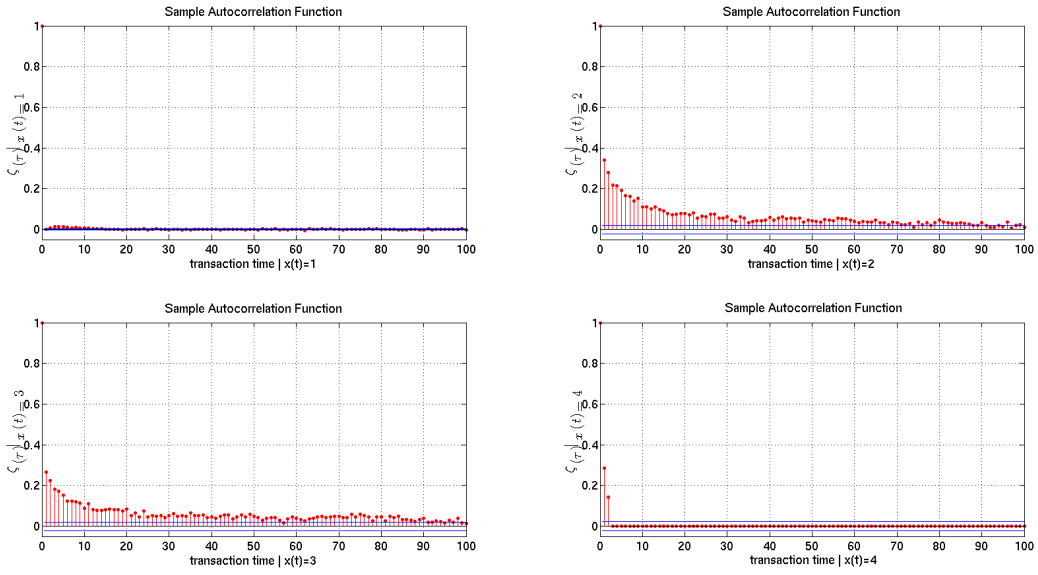

Figure 4.17: Empirical autocorrelation of price changes conditioned to specific values of transitions, i.e. $corr\left(\Delta p\left(t\right),\Delta p\left(t+\tau\right)|x\left(t\right)=i\right),\ i\in\{1,2,3,4\}$, for MSFT in the low regime.

correlation between price changes is significant. Such correlations imply the presence of inefficiencies in the mid-price changes process connected to spread transitions. The inefficiency seems to be due to the change in the spread value during the transition. The Markov models defined by the eq. 4.9 and eq. 4.44 imply that

$$corr\left(\Delta p\left(t\right),\Delta p\left(t+\tau\right)|x\left(t\right)=i\right)=0,\ \forall i\in\{1,2,3,4\}\,. \tag{4.45}$$

In this case, indeed, the price changes conditioned on the value of $x\left(t\right)$ are i.i.d variables. In the context of the DCMM(e,p) model defined by eq. 4.15, it is allowed the presence of correlations between price changes if we allow for $e\neq 0$. We do not explore this case in our work, but the empirical results of this section can motivate further investigations of the DCMM model in order to study the presence of inefficiencies/frictions in the mid-price process due to the coupled spread dynamics.

Chapter 5

Multiscale model selection by means of the Jensen–Shannon metric

Modeling financial time series at different time scales is still an open challenge. The choice of a suitable indicator quantifying the distance between the model and the data is therefore of fundamental importance for selecting models. In this chapter, we propose a multiscale model selection method based on the Jensen–Shannon distance in order to select the model that is able to better reproduce the distribution of price changes at different time scales. Specifically, we consider the problem of modeling the high frequency dynamics of two large tick stocks, i.e. MSFT and CSCO. We study the price process at different time scales and compute the Jensen–Shannon distance between the original dataset and the synthetic data simulated by means of models developed in chapter 4. This study shows that the coupling between spread and returns is important to model the return distribution at different time scales of observation, ranging from the scale of single transactions to the daily time scale.

5.1 Introduction

A specific challenge in financial modeling is how the return distribution changes at different time scales [118]. Due to the presence of fat-tailed distributions, also at very short time scales, and non-linear time correlations, the dynamics of returns distribution is far from trivial and not well described by any model [46]. The problem becomes even more dramatic when one wants to describe the price-change distribution also at the shortest time scales, i.e. when the discrete nature of trading appears. Trading and, correspondingly, price changes occur at discrete time. One of the main methodological problem is, therefore, to have a method to compare data and model predictions at different time scales.

In this work, we propose to perform multiscale model selection for financial time series by using the Jensen–Shannon distance [92, 68, 167]. We perform the model selection at different scales m, representing the level of aggregation of the time series.

In other words, given the price change time series, $x(t)$, we study the properties of the probability distribution of its sums $y_m = \sum_{t=1}^{m} x(t)$. It is important to clarify that we do not perform a goodness-of-fit test at different scales m defining a p-value relative to a specific statistic, e.g., Kolmogorov–Smirnov statistic [183], etc.. Our analysis consists, instead, in the comparison between the probability distribution computed from empirical data and those computed from synthetic data generated by specific statistical models. The discrepancy is measured by the Jensen–Shannon distance. We show that models containing the coupling between price and spread, as well as the time correlation of spread outperform other models without these characteristics in describing the change of the shape of the return distribution across scales.

The chapter is organized as follows. In Section 5.2, we illustrate the definitions of the Jensen–Shannon divergence and distance, and we characterize the unavoidable bias, due to the finiteness of the data sample. In Section 5.3, we apply the Jensen–Shannon distance criteria to select among four competing models of the dynamics of the price of a large tick stocks.

5.2 Jensen–Shannon Distance

Distance or divergence measures are of key importance in a number of theoretical and applied statistical inference and data processing problems, such as estimation, detection, compression and model selection [29]. Among the proposed measures, one of the best known is the Kullback–Leibler (KL) divergence between two distributions, $D(\mathbf{p}||\mathbf{q})$ [74], also called *relative entropy*. It is a measure of the inefficiency of assuming that the distribution is $\mathbf{q}$ when the true distribution is $\mathbf{p}$. It is used in many different applications, such as econometrics [230], clustering analysis [85], multivariate analysis [72, 239]. We will limit the following discussion to discrete probability distributions.

Let X be a discrete random variable with support of definition $\mathcal{X}$ and probability mass function $p(x)$, $x \in \mathcal{X}$. If $q(x)$ is another probability mass function defined on the same support, $\mathcal{X}$, the KL-divergence is defined as:

$$D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \tag{5.1}$$

where the base of the logarithm is two. We use the convention that $0 \log(0/0) = 0$ and the convention, based on continuity arguments, that $0 \log(0/q) = 0$. If there is any symbol, $x \in \mathcal{X}$, such that $p(x) > 0$ and $q(x) = 0$, then $D_{KL}(\mathbf{p}||\mathbf{q})$ is undefined. This means that distribution $\mathbf{p}$ has to be absolutely continuous with respect to $\mathbf{q}$ for KL-divergence to be defined [152]. It is well known that $D_{KL}(\mathbf{p}||\mathbf{q})$ is non-negative and additive, but not symmetric [152]. In order to overcome this problems, Lin [167] defined a new symmetric divergence, called L divergence:

$$D_L(\mathbf{p}, \mathbf{q}) = D_{KL}(\mathbf{p}||\mathbf{m}) + D_{KL}(\mathbf{q}||\mathbf{m}) \tag{5.2}$$

where $\mathbf{m} = (\mathbf{p} + \mathbf{q})/2$ is the "mean "probability mass function. $D_L(\mathbf{p}, \mathbf{q})$ vanishes if and only if $\mathbf{p} = \mathbf{q}$. The L divergence is symmetric and bounded by $D_L(\mathbf{p}, \mathbf{q}) \leq 2$. It is worth noticing that the L divergence can be expressed in terms of the Shannon entropy as:

$$D_L(\mathbf{p}, \mathbf{q}) = 2H\left(\frac{\mathbf{p} + \mathbf{q}}{2}\right) - H(\mathbf{p}) - H(\mathbf{q}) \tag{5.3}$$

i.e., it is the difference of entropy between the mean distribution, $\mathbf{m}$, and the sum of the entropies of $\mathbf{p}$ and $\mathbf{q}$. The generalization of the L divergence is the Jensen–Shannon divergence [167], defined as:

$$Div_{JS}(\mathbf{p}, \mathbf{q}) = H(\pi_1\mathbf{p} + \pi_2\mathbf{q}) - \pi_1 H(\mathbf{p}) - \pi_2 H(\mathbf{q}) \tag{5.4}$$

where $\pi_1, \pi_2 \geq 0$, $\pi_1 + \pi_2 = 1$ are the weights of the probability distributions, $\mathbf{p}$ and $\mathbf{q}$, respectively. According to this new definition, $D_L(\mathbf{p}, \mathbf{q}) = 2Div_{JS}(\mathbf{p}, \mathbf{q})$, for $\pi_1 = \pi_2 = 1/2$. Endres *et al.* [92] found that the square root of D_L is a metric, *i.e.*, it fulfills the triangle inequality. They named this new information metric the Jensen–Shannon distance, $\mathcal{D}_{JS}$:

$$\mathcal{D}_{JS}(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{x \in \mathcal{X}} \left(p(x) \log \frac{2\,p(x)}{p(x) + q(x)} + q(x) \log \frac{2\,q(x)}{p(x) + q(x)} \right)} \tag{5.5}$$

The bounds of this distance are: $0 \leq \mathcal{D}_{JS} \leq \sqrt{2}$. The Jensen–Shannon divergence is used also in statistical mechanics [77], thermodynamics [76], networks [56], particle physics [59], biology [98].

In this chapter, we are interested in using the Jensen–Shannon distance as a method for selecting among a set of models the one that best describes a given dataset. We are concerned with the case when our data is represented by a discrete time series of length N. When considering different competing models, we search for the best model describing the probability distribution of the aggregation (*i.e.*, sum) of the time series at different time scales m. Moreover, the use of Jensen–Shannon distance allows us to compare two empirical distributions.

To be more specific, consider the random variable, x, taking values from the set $\mathbf{x} = (x_1, \cdots, x_k)$ with probabilities $\mathbf{p} = (p_1, \cdots, p_k)$. Given N observations of the time series, $x(t)$, $t = 1, .., N$, one builds a histogram $\mathbf{n} = (n_1, \cdots, n_k)$, where n_i is the number of times the outcome was x_i. The frequency vector $\mathbf{f} = (f_1, \cdots, f_k) = (n_1/N, \cdots, n_k/N)$ is an estimator of the probability distribution, $\mathbf{p}$. We want to perform a statistical analysis at different scales of aggregation, *i.e.*, we study the probability distribution, $\mathbf{p}_m$, and frequency distribution, $\mathbf{f}_m$, of the sum $\sum_{t=1}^{m} x(t)$, where the value, m, defines the scale. The probability distribution of the elementary process $x(t)$, corresponding to $m = 1$, is denoted by $\mathbf{p}_{m=1} = \mathbf{p}$. If the initial dataset had N values, the scale, m, is limited by $1 \leq m \leq N$. The number of experimental data available at each aggregation scale m reduces to $N_m \equiv \lfloor N/m \rfloor$, because we sum the experimental data, which belong to the N_m non-overlapping windows of length m.

In order to select the best model that describes the data at all aggregation scales, we compute the Jensen–Shannon distances for various values of m, *i.e.*, $\mathcal{D}_{JS}\left(\mathbf{p}_m, \mathbf{f}_m\right)$. We estimate $\mathbf{p}_m$ according to different statistical models, and we select the one that minimizes $\mathcal{D}_{JS}\left(\mathbf{p}_m, \mathbf{f}_m\right)$ for the different values of m. As will be clear below, we will also need to compute the distance between two frequency distributions in order to compare the two different datasets, $\mathcal{D}_{JS}\left(\mathbf{f}_{1,m}, \mathbf{f}_{2,m}\right)$. In this case, we assume that the length of the two datasets is the same $N_1 = N_2 = N$.

It is important to stress that even if we knew the *true* distribution, $\mathbf{p}_m$, the distance, $\mathcal{D}_{JS}\left(\mathbf{p}_m, \mathbf{f}_m\right)$, inferred from a finite sample of data, would be larger than zero. The fluctuations of $\mathbf{f}_m$ from dataset to dataset may not only result in fluctuations of the numerical values of $\mathcal{D}_{JS}$, but also in a systematic shift, *i.e.*, bias, of the numerical values of $\mathcal{D}_{JS}$. This bias is identified with the expectation value, $E\left[\mathcal{D}_{JS}\left(\mathbf{p}_m, \mathbf{f}_m\right)\right] \neq 0$, for the various values of the scale, m. The bias is also present if we compute the distance, $\mathcal{D}_{JS}\left(\mathbf{f}_{1,m}, \mathbf{f}_{2,m}\right)$, between two frequency vectors that are computed from datasets representing the same stochastic process.

The concept of a systematic bias of the numerical values of Jensen–Shannon divergence, Div_{JS}, is well known in the literature, and it is connected to the systematic bias in the estimation of entropy. It follows directly from Jensen inequality [74] that the expected value, $E\left[H\left(\mathbf{f}\right)\right]$, of the entropy computed from an ensemble of finite-length sequences cannot be greater than the theoretical value, $H\left(\mathbf{p}\right)$, of the entropy computed from the (unobservable) probabilities:

$$E\left[H\left(\mathbf{f}\right)\right] \leq H\left(\mathbf{p}\right) \tag{5.6}$$

where the expectation is defined over the ensemble of finite-length i.i.d.sequences generated by the probability distribution, $\mathbf{p}$. It can be shown that the expected value of the observed entropy is systematically biased downwards from the true entropy:

$$E\left[H\left(\mathbf{f}\right)\right] = H\left(\mathbf{p}\right) - \frac{k-1}{2N\ln 2} + O\left(N^{-2}\right) \tag{5.7}$$

where k is the number of components of the probability and frequency vectors, $\mathbf{p}$ and $\mathbf{f}$, and N is the ensemble size. This result was obtained by Basharin [28] and Herzel [138], who pointed out that to the first order, $O\left(1/N\right)$, the bias is independent of the actual distribution, $\mathbf{p}$. The term of order $O\left(1/N^2\right)$ involves the unknown probabilities $\mathbf{p} = \left(p_1, \cdots, p_k\right)$ and cannot be estimated in general [228, 218, 122].

Grosse *et al.* [124] derived an analytical approximation of the expected value of $Div_{JS}\left(\mathbf{f}_1, \mathbf{f}_2\right)$ between two i.i.d. sequences of length N coming from the same probability distribution, which is:

$$E\left[Div_{JS}\left(\mathbf{f}_1, \mathbf{f}_2\right)\right] = \frac{k-1}{4N\ln 2} + O\left(N^{-2}\right) \tag{5.8}$$

Clearly, also in the case of the Jensen–Shannon distance, $\mathcal{D}_{JS}$, there is a systematic positive bias.

5.2.1 A Simple Binomial Model

In this section, we present a toy example of the use of Jensen–Shannon distance for model selection. The purpose of the section is mostly didactical and serves to show the multiscale procedure and the issues related to the finiteness of the sample that will be present also in the real financial case described in the next section.

Let us consider a process, which at scale $m = 1$ is a binomial i.i.d. process, *i.e.*, $\mathbf{p}_{m=1}$ is described by $B(n, p_B)$, where p_B describes the probability of success. The sum of m i.i.d. binomial variables is still described by a binomial distribution [145], *i.e.*, $\mathbf{p}_m = (p_{m,1}, \cdots, p_{m,k})$ is described by $B(nm, p_B)$, and its support is a set composed by $k = nm + 1$ elements. Given a time series of length N, at each aggregation scale, m, we have $N_m \equiv \lfloor N/m \rfloor$ observations from non-overlapping windows, and we measure the frequency vector $\mathbf{f}_m = (n_{m,1}/N_m, \cdots, n_{m,k}/N_m)$, where $n_{m,i}$ is the number of occurrences of the event, i, at scale m.

The probability distribution of empirical frequencies is given by the multinomial distribution:

$$p(\mathbf{f}_m) = \lfloor N/m \rfloor ! \prod_{i=1}^{nm+1} \frac{p_{m,i}^{n_{m,i}}}{n_{m,i}!} \tag{5.9}$$

In principle, one can compute exactly the moments of the distances, $\mathcal{D}_{JS}(\mathbf{p}_m, \mathbf{f}_m)$ and $\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2)$, which are:

$$E\left[\mathcal{D}_{JS}^b(\mathbf{p}_m, \mathbf{f}_m)\right] = \sum_{n_{m,1},\cdots,n_{m,k}=0}^{\lfloor N/m \rfloor} \lfloor N/m \rfloor ! \prod_{i=1}^k \frac{p_{m,i}^{n_{m,i}}}{n_{m,i}!} \left(\sum_{i=1}^k \left(f_{m,i} \log\left(\frac{2f_{m,i}}{f_{m,i}+p_{m,i}}\right) + p_{m,i} \log\left(\frac{2p_{m,i}}{f_{m,i}+p_{m,i}}\right)\right)\right)^{b/2} \tag{5.10}$$

and

$$E\left[\mathcal{D}_{JS}^b\left(\mathbf{f}_m^1, \mathbf{f}_m^2\right)\right] = \sum_{n_{m,1}^1,\cdots,n_{m,k}^1=0}^{\lfloor N/m \rfloor} \sum_{n_{m,1}^2,\cdots,n_{m,k}^2=0}^{\lfloor N/m \rfloor} \left(\lfloor N/m \rfloor !\right)^2 \prod_{j=1}^{\lfloor N/m \rfloor} \prod_{l=1}^{\lfloor N/m \rfloor} \frac{\left(p_{m,j}^1\right)^{n_{m,j}^1}}{n_{m,j}^1!} \frac{\left(p_{m,l}^2\right)^{n_{m,l}^2}}{n_{m,l}^2!}$$
$$\left(\sum_{i=1}^k \left(f_{m,i}^1 \log\left(\frac{2f_{m,i}^1}{f_{m,i}^1+f_{m,i}^2}\right) + f_{m,i}^2 \log\left(\frac{2f_{m,i}^2}{f_{m,i}^1+f_{m,i}^2}\right)\right)\right)^{b/2} \tag{5.11}$$

These expressions can be used to compute the mean and variance of the Jensen–Shannon distance, as well as of the Jensen–Shannon divergence.

The computational problem with these expectations are the values of $k = nm + 1$ and of N, because the number of categories of the multinomial distribution grows dramatically with the scale, m. The support of the multinomial distribution for the scale, m, has a number of elements:

$$n.e. = \binom{\lfloor N/m \rfloor + nm}{nm} = \frac{\prod_{i=1}^{nm} \left(\lfloor N/m \rfloor + i\right)}{(nm)!} \tag{5.12}$$

For example if $N = 1000$ and $n = 2, m = 1$, we have that the number of elements is $n.e. \approx 5 \times 10^5$.

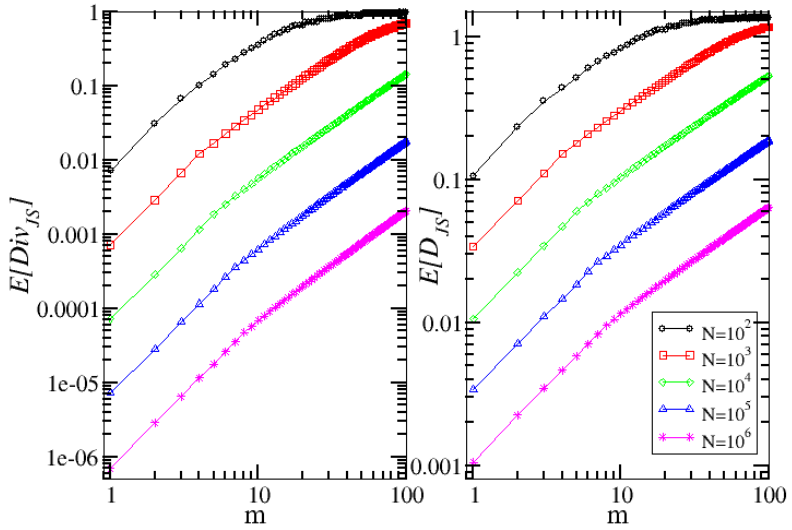

Figure 5.1: $E\left[Div_{JS}\left(\mathbf{f}_m^1, \mathbf{f}_m^2; N\right)\right]$ (**left**) and $E\left[\mathcal{D}_{JS}\left(\mathbf{f}_m^1, \mathbf{f}_m^2; N\right)\right]$ (**right**) for the binomial model as a function of the aggregation scale, m, and for different values of time series length, N. Results are obtained from numerical simulations, and the plots are in log-log scale.

To handle this problem, we compute these expectations by means of Monte Carlo simulations, and we replace ensemble averages with sample averages, *i.e.*, for example:

$$E\left[\mathcal{D}_{JS}\left(\mathbf{p}_m, \mathbf{f}_m\right)\right] \approx \left\langle \mathcal{D}_{JS}\left(\mathbf{p}_m, \mathbf{f}_m\right)\right\rangle \equiv \frac{1}{N_r}\sum_{j=1}^{N_r} \mathcal{D}_{JS}\left(\mathbf{p}_m, \mathbf{f}_m\left(\mathbf{n}_{m,j}\right)\right) \tag{5.13}$$

where $\langle \cdots \rangle$ represents the ensemble average and j represents the j-th simulation of the set of N_r.

We first consider the problem of the finite sample bias in the computation of the Jensen–Shannon divergence and distance. Specifically, we compute $E\left[\mathcal{D}_{JS}\left(\mathbf{f}_m^1, \mathbf{f}_m^2; N\right)\right]$ and $E\left[Div_{JS}\left(\mathbf{f}_m^1, \mathbf{f}_m^2; N\right)\right]$ as a function of the time series length, N, when the two frequency vectors are taken by two independent realizations of the same binomial model. We study the two information functionals in the range $m = [1, \cdots, 100]$ for the values $N = 10^2, 10^3, 10^4, 10^5, 10^6$, as reported in Figure 5.1.

As expected, the bias decreases with N and increases with m. By using the result in Equation (5.8) for sequences of i.i.d observations, we are able to compute analytically the shape of the initial part of the curve corresponding to the Jensen–Shannon divergence. In fact, in our framework, we should perform the following substitutions in Equation (5.8), $N \to N/m$ and $k \to nm + 1$, and we thus obtain that the scaling of the Jensen–Shannon divergence as a function of N and m for the

binomial model is:

$$E\left[Div_{JS}\left(\mathbf{f}_m^1, \mathbf{f}_m^2; N\right)\right] = \frac{nm^2}{4N\ln(2)} + O\left(\frac{m^2}{N^2}\right) \tag{5.14}$$

This approximation is more and more valid when N increases, as we can observe in Figure 5.1. A power-law fit $f(x) = cx^e$ on the initial part of the curves in the case $N = 10^6$ gives $c = (7.1 \pm 0.1) \times 10^{-7}, e = (ch3eq : 2.0 \pm 0.1)$. This is in agreement with the power law of exponent 2 of Equation (5.14) and with the coefficient $n/(4N\ln(2)) \approx 7.2 \times 10^{-7}$.

In the case of the Jensen–Shannon distance, we do not have any analytic result and limit ourselves to a power-law fit of the initial part of the curve. For the case $N = 10^6$, the fit gives $c = (1.1 \pm 0.1) \times 10^{-3}, e = (1.0 \pm 0.1)$. The initial part of the curve appears to scale linearly with scale m, *i.e.*, $E\left[Div_{JS}\left(\mathbf{f}_m^1, \mathbf{f}_m^2; N\right)\right] \propto m$.

In order to illustrate how to perform model selection with the Jensen–Shannon distance, we consider the case of an (artificial) sample generated from the binomial model with $p_B = 0.5$. We then compare the Jensen–Shannon distance between this sample and another realization of the model with the same parameter and of a realization of the model with different parameter $p_B \neq 0.5$. As expected, Figure 5.2 shows that the expected value of the Jensen–Shannon distance between two samples generated by the model with the same parameter is always smaller than the distance between two samples with a different parameter. Moreover, the distance between a sample and the true probability distribution is smaller than the distance between two samples of the same model. This simple observation suggests to us a procedure for selecting models by using the Jensen–Shannon distance.

Specifically, suppose that $\mathbf{f}_{m=1}^{sam}$ represents the frequency vector computed from the sample of length N, but we do not known the true model that generates it. Suppose we have a statistical model, from which we are able to simulate an output of the same length. In this case, we can compute a frequency, $\mathbf{f}_{m=1}^{mod}$, from our reference model. To compare the two processes at different scales m, we compute the frequencies, $\mathbf{f}_m^{sam}$ and $\mathbf{f}_m^{mod}$, from the sums of the initial sample over $\lfloor N/m \rfloor$ non-overlapping windows. If we have different competing models $M_1, M_2, \cdots$, we generate synthetic samples of length N and compute the distances $\mathcal{D}_{JS}\left(\mathbf{f}_m^{sam}, \mathbf{f}_m^{mod}; l\right)$, where the index, l, runs on the possible different models. The model that minimizes the Jensen–Shannon distance at different scales m is the model that reproduces the data better. It is clear that even if we had the true model, the minimum distance at different scales will be different from zero. This is because, as we have seen before, $E\left[D_{JS}\left(\mathbf{f}_m^1, \mathbf{f}_m^2; N\right)\right]$ is larger than zero, even when the two samples come from the real model. As we will see in the financial case in the next section, one can split the real sample into two subsamples of length $N/2$ and compute their Jensen–Shannon distance, to be used as a reference line with respect to the Jensen–Shannon distance between the data and the models.

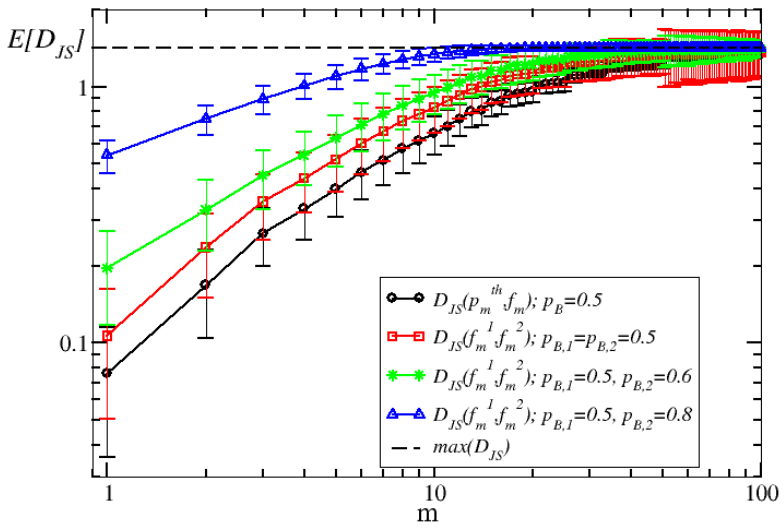

Figure 5.2: Expectations and standard deviations of the Jensen–Shannon distance between two samples of the binomial model with the same parameter $p_{B,1} = p_{B,2} = 0.5$ (red squares) and with different parameters (green diamonds and blue triangles). The black circles are an estimation of the Jensen–Shannon distance between a sample and the true model.

5.3 Application to High Frequency Financial Data

In this section, we use the above multiscale procedure, based on the Jensen-Shannon distance, in order to select the best statistical model in the particular case of models describing the high frequency price dynamics of large tick stocks developed in chapter 4. Our purpose is to show the importance of the spread coupling in order to reproduce price change distributions.

We want to compare the three models introduced in section 4.3, i.e the MS_B, MS and DCMM(50) models. In this section we introduce also a basic model, i.e. that does not allow for spread couplings, in order to highlight the importance of the spread coupling in order to reproduce the entropic information content of price changes distribution. The model $M0$ is defined by an i.i.d process for $x(t)$, where the unconditional distribution, $p(x(t))$, reproduces the empirical distribution of price changes after one transaction. As we have observed in section 4.9.3, this implies a small asymmetry between positive and negative values of the price change distribution. In this case, $p(x(t)|s(t)) = p(x(t))$, i.e. we have independence between the two variables x and s.

The problem is how to select the model that reproduces the data better. We focus our selection problem on the ability of the models to reproduce the price-change

process at different time scales. The selection problem does not involve the spread process, $s(t)$.

First, we compute the Jensen–Shannon distance between two realizations of the real process. To this end, we divide the sample into two non-overlapping samples, each of length $N/2$, and we compute the two frequency vectors, $\mathbf{f}_m^1$ and $\mathbf{f}_m^2$, for each value of m. We then compute the Jensen–Shannon distance, $\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N/2)$. It is clear that this is only one of the possible values of the random variable, $\mathcal{D}_{JS}$, and we expect that it will be affected by some kind of fluctuations. Then, we generate $N_r = 25$ synthetic samples of length $N/2$ of processes corresponding to our four models. In this way, we compute N_r different frequencies $\mathbf{f}_m^{model}$ that allow us to compute the sample averages:

$$\langle \mathcal{D}_{JS}^q \left(\mathbf{f}_m^{sample}, \mathbf{f}_m^{model}; N/2\right)\rangle \;=\; \frac{1}{N_r}\sum_{i=1}^{N_r} \mathcal{D}_{JS}^q \left(\mathbf{f}_m^{sample}, \mathbf{f}_m^{model,i}; N/2\right) \qquad (5.15)$$

from which we can compute a mean and a standard deviation value for the Jensen–Shannon distance for each value of m. The results for the different models are reported in figures 5.3,5.4,5.5,5.6 for the four time series under analysis. The model that reproduces the empirical data better, i.e. which is closer to $\mathcal{D}_{JS}(\mathbf{f}_m^1, \mathbf{f}_m^2; N/2)$, is the DCMM model. We observe that the difference between the distances computed from the MS and DCMM Montecarlo simulations are almost absent. This implies that the MS model is sufficient in order to reproduce the price changes distribution. The improvement given by the DCMM model is given by the presence of autocorrelations between squared price changes and not by his ability to reproduce price changes distributions. It is important to notice that also the MS_B model reproduces the empirical data for values of the scale $m > 10$. In fact, for $m > 10$, the MS and MS_B models have the same ability to reproduce the empirical data. The conditioning rules of Equation 4.9 are critical in order to reproduce the data for values of $m > 10$. The model $M0$, instead, appears to reproduce the data better for the scale of a single transaction, *i.e.*, $m = 1$. This can be only the consequence of the fact that the probability distribution of $M0$ models is exactly the same as the empirical distribution of price changes for single transactions, *i.e.*, it reproduces the small asymmetry of the real distribution between positive and negative values of price changes. Instead, MS and MS_B have symmetric distributions for price changes for single transactions. We can observe that the three models reproduces the data for $m > 3,000$ equally well. This corresponds to a real time of the order of one hour, *i.e.*, the daily time scale. This time scale can be interpreted as the one after which the microscopic details of price formation and market microstructure are not relevant anymore in describing the dynamics of the shape of the price change distribution. In other words, the coupling of the price-change process and of the bid-ask spread process appears to be the key to understand the dynamics of prices for a large tick stock from the time scale of single transaction to the daily time scale.

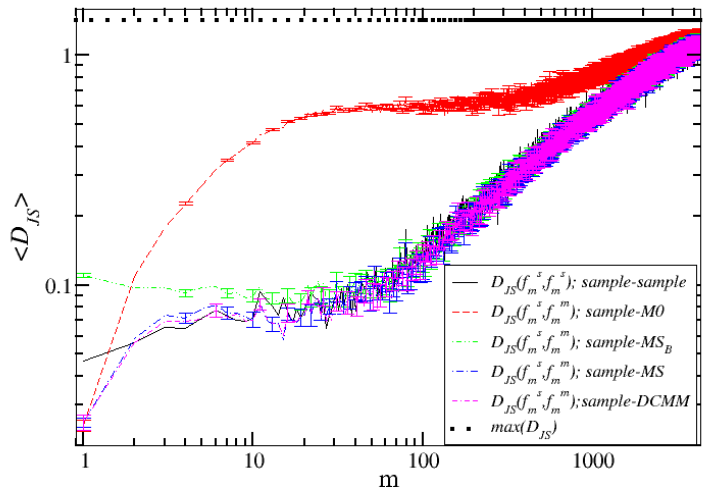

Figure 5.3: Mean and standard deviation of $\mathcal{D}_{JS}$ between Microsoft(LOW) data and three models, namely $M0$, MS and MS_B (see the text). The black line is the distance, $\mathcal{D}_{JS}$, between the two subsamples of the real data obtained by splitting the sample in two. We do not display the error bars for each value of m, but only for 25% of them.

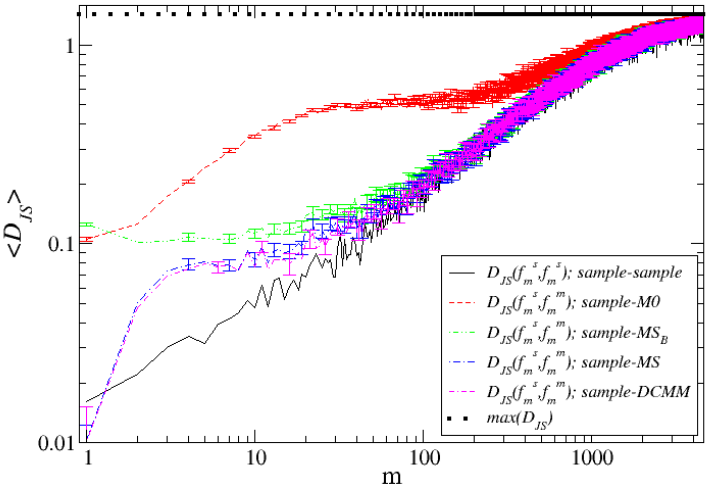

Figure 5.4: Mean and standard deviation of $\mathcal{D}_{JS}$ between Microsoft(HIGH) data and three models, namely $M0$, MS and MS_B (see the text). The black line is the distance, $\mathcal{D}_{JS}$, between the two subsamples of the real data obtained by splitting the sample in two. We do not display the error bars for each value of m, but only for 25% of them.

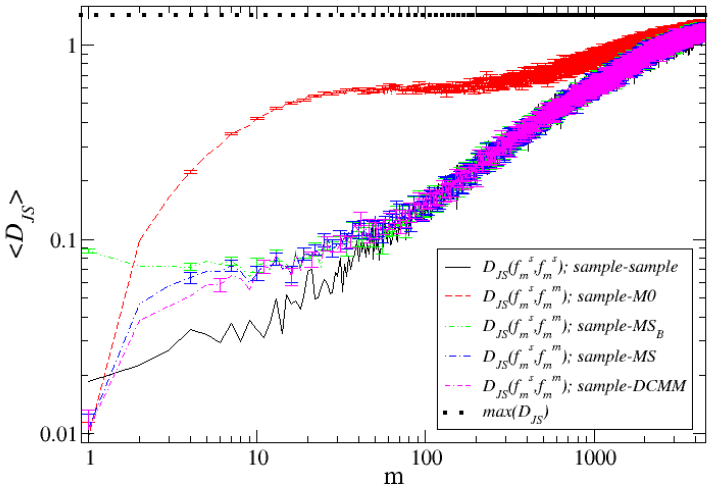

Figure 5.5: Mean and standard deviation of $\mathcal{D}_{JS}$ between Cisco(LOW) data and three models, namely $M0$, MS and MS_B (see the text). The black line is the distance, $\mathcal{D}_{JS}$, between the two subsamples of the real data obtained by splitting the sample in two. We do not display the error bars for each value of m, but only for 25% of them.

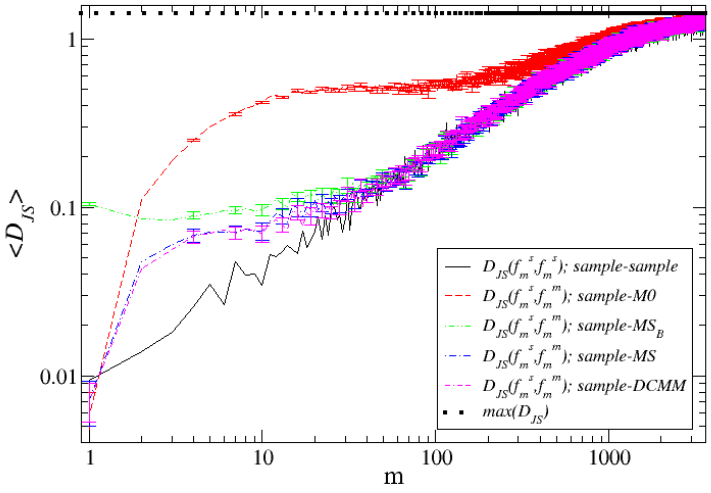

Figure 5.6: Mean and standard deviation of $\mathcal{D}_{JS}$ between Cisco(HIGH) data and three models, namely $M0$, MS and MS_B (see the text). The black line is the distance, $\mathcal{D}_{JS}$, between the two subsamples of the real data obtained by splitting the sample in two. We do not display the error bars for each value of m, but only for 25% of them.

5.4 Appendix: Kolmogorov-Smirnov test

In this section we use a classic non-parametric test, i.e. the Kolmogorov-Smirnov two sample test [183], in order to test if two empirical distributions came from the same distribution. In this way it is possible to test the goodness of the above models independently from the procedure defined by the JS-distance proposed in this chapter. We follow the same procedure of section 5.3, i.e. we divide the sample coming from time series in two and compute the KS-test between the two sub-samples. Such procedure is repeated between one sub-sample and data coming from Montecarlo simulations of the $M0$, MS_B and MS models. The errors on such procedure are obtained repeating the KS-test on $N_r = 25$ different Montecarlo simulations of the same process. We do not include the DCMM model because its performance are almost the same given by MS model in order to reproduce price changes distributions. We illustrate only the results regarding the MSFT stock during the low volatility regime, the other time series show the same behavior of MSFT low.

The Kolmogorov-Smirnov test is based on the concept of the *empirical* c.d.f.. Let us denote $F(\bar{x}) = p(x < \bar{x})$ a c.d.f. of a true underlying distribution of the data. If N is the length of the data sample, the empirical distribution function is defined by

$$F_N(\bar{x}) = \frac{1}{N} \sum_{i=1}^{N} I(x < \bar{x}) \tag{5.16}$$

where I is the indicator function. For a two sample test we compute two empirical c.d.f. $F_{N_1}(\bar{x})$ and $G_{N_2}(\bar{x})$ and we want to test the following hypothesis

$$H_0 : F = G \;\; vs \;\; F \neq G, \tag{5.17}$$

were the F and G are the two underlying exact c.d.f.. To perform the test one compute the following statistic

$$D_{N_1 N_2} = \sup_{\bar{x}} |F_{N_1}(\bar{x}) - G_{N_2}(\bar{x})|, \tag{5.18}$$

and compute a critical value $D_{N_1 N_2,\alpha}$[1] at a significance level α. We reject the null hypothesis if $D_{N_1 N_2} > D_{N_1 N_2,\alpha}$. Such statistic fundamentally computes the maximum distance between the two empirical distributions. In our case $N_1 = N_2 = N/2$. We perform such test at different time scales m, i.e. we compute $F_{N_1,m}$ and $G_{N_2,m}$. Such analysis is showed in fig. 5.7. The graphs show the two most used levels of significance, i.e. 0.05 and 0.01. If the statistic is lower than the chosen level, we reject the H_0 null hypothesis. The results clearly show an increasing ability of the proposed models to reproduce the empirical distributions at different aggregation time scales. The KS-test confirms that the MS model is the best model, because it is able to reproduce the empirical distribution for time scales $m > 6$.

[1]For N_1 and N_2 sufficiently large, $D_{N_1 N_2,\alpha} = c(\alpha)\left(\frac{N_1+N_2}{N_1 N_2}\right)^{1/2}$, where $c(\alpha)$ is the inverse of the Kolmogorov distribution at α.

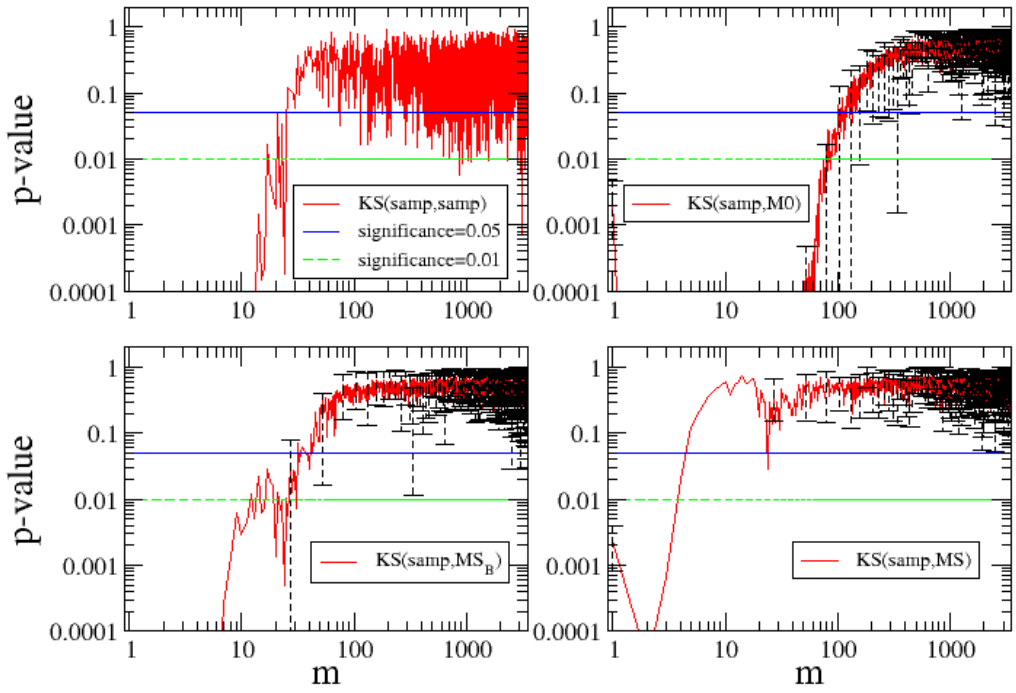

Figure 5.7: P-values coming from the Kolmogorov-Smirnov two sample test. The first graph in the upper-left corner shows p-values for two sub-samples of the original time series. The other three graphs show the KS-test performed using always the same sub-sample of the real data, i.e. MSFT low, and 25 different Montecarlo simulations of $M0$, MS_B and MS models. We report two different significance levels, i.e. 0.05 and 0.01. We do not display the error bars for each value of m, but only for 25% of them.

One possible critique to the use of a KS-test in the present case is that the results concerning the KS-test are valid for a continuous random variables, instead the price changes are discrete random variables. It is known from Noether et al. [193, 115] that the KS-test is conservative if $F(\bar{x})$ is discrete. This means it makes more difficult to reject the null hypothesis, i.e. p-values from a conservative test tend to be too large. Despite this behavior, it is more important the ranking of models than the effective value of the time scale m after which we cannot reject the null hypothesis. We conclude that the KS-test and the information-based JS-distance give the same results, i.e. the latent spread coupling is fundamental in order to reproduce the price change distributions at different time scales by means of the MS model.

Part II

Market impact and optimal strategies

Chapter 6

An introduction to market impact and optimal execution

The second part of the thesis work, i.e. chapters 7 and 8, deals with the analysis of optimal execution strategies in the context of a nonlinear transient market impact model [43, 104]. In this chapter we illustrate the fundamental concept of market impact of trades on asset's prices in section 6.1. In section 6.2 we review the optimal execution problem and the definitions of price manipulations. We report the constraints imposed by the absence of price manipulations on the Almgren-Chriss model [15, 16] and on the propagator model [43, 104].

6.1 Market impact

The relation between the transacted volume and the consequent expected price shift is called the *market impact* function. We now discuss the impact of individual transactions in limit order books markets, whose volume we will denote by q. Many studies have examined the market impact for a single transaction, and all have observed a concave function of the transaction volume, that is, one that increases rapidly for small q and more slowly for larger q. Lillo et al. [165] analyzed the 1,000 largest firms on the NYSE during the period 1995-1998 recorded on TAQ (Traded and Quote) database. They studied roughly 113 million transactions and 173 million quotes, investigating the shift in the midprice caused by the most recent transaction. The expected price shift is well fitted by a concave power law function

$$\mathbb{E}\left[r\left(t_i\right)|q\left(t_i\right)\right] = \frac{sign\left(q\right)\ |q|^{\delta}}{\lambda}, \tag{6.1}$$

where $sign\left(q\right) = 1$ represents a buy trade and $sign\left(q\right) = -1$ a sell trade, and λ is a liquidity parameter that depends on market capitalization. The exponent $\delta = \delta\left(q\right)$ is approximately 0.5 for small volumes and 0.2 for large volumes. For the purpose of this work we are more interested to stress the functional form of the impact, than focus on details like the liquidity parameter λ. Bouchaud et al. [43] observed that a logarithmic function gave the best fit in the case of French stocks. Other empirical

studies, focused on the dependence of impact function on volumes, confirmed that this function is a strongly concave function of the volume [132, 147, 204, 209].

The market impact is one of the main components of the cost of trading, particularly for large firms that execute large orders, i.e. the so called *metaorders* [42]. Such orders are typically split into small pieces and executed incrementally. Other cost can be represented by commissions and fees, that are explicitly stated and easily measured, or by market frictions such as the bid-ask spread.

6.1.1 Metaorder impact

The metaorders are large institutional orders and are also known as hidden orders. They are hidden in the sense that traders keep the true order size secret to minimize information leakage. Rather than being traded as a block, these orders are split into small slices, i.e. the *child orders*, and executed incrementally. The strategic reason for the incremental execution of metaorders were originally analyzed by Kyle [153], who developed a model for an insider trader with monopolistic information about future prices. Kyle showed that the optimal strategy for the trader is to break his metaorder into pieces and execute it incrementally at a uniform rate, gradually incorporating his information into price.

The study of the market impact of metaorders requires a different approach compared to individual or aggregate orders [237, 34, 87, 190, 238, 184, 254, 26]. The reason is that metaorders come from the same agent and generate strong correlations in order flow trough a series of incremental executions [166]. This literature regards mainly two different issues. The first issue is to study the expected price shift between the beginning and the end of a metaorder of size X. The second issue is to study how the price changes during the execution of the metaorder and the price dynamics after the completion of the execution. One well known result regarding the first issue is the *square-root* law [237]. More precisely, the expected return between the first and last trade of a metaorder of size X is empirically described by

$$r\left(X\right) = Y\,\sigma\,\sqrt{\frac{X}{V}}, \tag{6.2}$$

where σ is the daily volatility of the asset, V is the total daily traded volume, and the numerical constant Y is of order unity. The existing empirical studies substantially agree with this power law functional form [190, 238, 34, 26, 17, 184]. More recently Zarinelli et al. [254] observed that a logarithmic functional form can describe data better than the power-law (square-root) functional form. They observe that the square root is an approximation over part of the metaorder volume range, i.e. the square root does not fit well data representing metaorders of small volumes.

The empirical results regarding the second issue leads to the conclusion that as a given metaorder is executed the impact grows in time according to a power-law, after the metaorder is executed, the price reverts to a level of about $0.5 - 0.7$ of its value at its peak. Moro et al. [190], studying the LSE (London Stock Exchange) and BME (Bolsas Mercados Españoles), found an average impact during the execution of

an order during the time interval $[0, T]$ described by the following functional form

$$\begin{aligned} r(t) &\approx (4.28 \pm 0.21) \times \left(\frac{t}{T}\right)^{(0.71\pm0.03)}, (BME) \\ r(t) &\approx (2.13 \pm 0.05) \times \left(\frac{t}{T}\right)^{(0.62\pm0.02)}, (LSE). \end{aligned} \tag{6.3}$$

After the completion of the metaorder the price reverts to a *permanent* value. Bershova and Raklin [34] observed the relaxation dynamics of the impact for US equity market. They found that the impact decay is a multi-regime process, approximated by a power law in the first few minutes after order completion and subsequently by exponential decay.

Different models attempted to explain the shape of metaorder market impact, i.e. the dependence on the its total volume X and on its duration. Farmer et al. [97] proposed a theory based on the traditional view in finance that market impact reflects information. They derive a fair pricing principle that, when combined with the martingale property of prices, predicts that the average execution price should equal the final price when the metaorder has completed and prices have been allowed to relax. This model can be viewed as an extension of Kyle original model [153], but with more realistic assumptions. According to this theory the market impact should increase as the square root of the metaorder size, with an average market impact relaxing to about two-thirds of peak impact. Toth et al. [238] present a theory for market impact based on the concept of a latent order book. The key idea is that the true order book does not reflect the actual supply and demand that are present in the market, due to the fact that participants do not reveal their true intensions. They show that for prices to be diffusive, i.e. for the variance to grow linearly with time, it is necessary for the latent order book to have a linear profile around the current price, which implies a square root impact function. This is supported by simulations of a simple agent-based model. In a recent work Brokmann et al. [50] study specifically the metaorder impact decay profile after completion. They highlight a methodological problem with the way data has been analyzed in all previous work on the subject. The point is that proprietary order flows tend to be highly auto-correlated in time due to the slicing of the metaorder during the day. Price dynamics seen after completion of an intra-day metaorder therefore reflects both its own decaying impact and the impact dynamics from all previous orders as well. Analyzing CFM (Capital Found Management) trades on equity market, after an appropriate deconvolution, they found that the *mechanical* impact of metaorders decays all the way to zero. This study shows that auto-correlated order-flows and trade information contents fully accounts for the apparent plateau observed in the raw data.

The short summary given so far shows that the literature presents two different point of view on the market impact. According to the first one [153, 134], more followed by the economic community, the market impact is the way information is conveyed to the market. In this framework, large investors react to new information by trying to successfully forecast price movements. Thus, they update their expectation

by using meta-orders, which leads to a new global equilibrium resulting in new price levels. According to this vision, meta-orders reveal the fundamental value of the price but do not really provoke it. According to the second one [42], more followed by mathematics and physics community, the market impact is created by the mechanical interaction between the single orders, i.e the child orders, and the limit order book, which leads to a long term imbalance in supply and demand.

6.2 Optimal execution

In this section we introduce the problem of the optimal execution and some mechanical impact models, according to which an optimal strategy can be determined. In this chapter and in chapters 7, 8 the continuous asset price will be referred to as $S(t)$.

The metaorders are usually fragmented in *child orders* in order to reduce the cost due to market impact. The optimization of trading strategies has long been an important goal for investors in financial markets [10]. As demonstrated in the context of a linear equilibrium model by Kyle almost thirty years ago [153], the optimal strategy for an investor with insider information on the fundamental price of an asset is to trade incrementally through time. This strategy allows the trader to minimize costs whilst also minimizing the revelation of information to the rest of the market. The precise way in which it is optimal to split the large order [35, 254] depends on the objective function and on the market impact model, *i.e.* the change in price conditioned on signed trade size. In part due to the increasing tendency toward a full automation of exchanges and in part due to the discovery of new statistical regularities of the microstructure of financial markets, the problem of optimal execution is receiving growing attention from the academic and practitioner communities [18, 108, 107, 194, 156, 10]. The particular way in which the execution of an order is scheduled can be critical, as is illustrated by the Flash Crash of May 6, 2010. According to CFTC-SEC (2010) [1], a important contribution in triggering this event was the extremely rapid execution of a larger order of certain futures contracts.

To generate order execution algorithms, one usually starts by setting up a stochastic market impact model that describes both the volatile price evolution of assets and how trades impact the market price as they are executed. One then specifies a cost criterion that can incorporate both the liquidity costs arising from market impact and the price risk resulting from late executions. Optimal trading trajectories, which are the basis for trading algorithms, are then obtained as minimizers of the cost criterion among all trading strategies that liquidate a given asset position within a given time frame. Some of such models admit an optimal order execution strategy. In others, an optimal strategy does not exist or shows unstable behavior.

6.2.1 Price manipulations

The phenomenon of price impact becomes relevant for orders that are large in comparison to the instantaneously available liquidity, i.e. number of shares, in markets. Such orders cannot be executed at once but need to be unwound over a certain time

interval $[0, T]$ by means of a dynamic order execution strategy. Such a strategy can be described by the asset position $x(t)$ held at time $t \in [0, T]$. The final position X is positive for a buy strategy and negative for a sell strategy. The condition $x(T) = X$ assures that the initial position has been unwound by time T. The path will be nonincreasing for a pure sell strategy and nondecreasing for a pure buy strategy. A general strategy may consist of both buy and sell trades and hence can be described as the sum of a nonincreasing and a nondecreasing strategy. That is, $x(t)$ is a path of finite variation.

A market impact model basically describes the quantitative effect of such an order execution strategy on asset prices. It usually starts by assuming exogenously given asset price dynamics $S_0 = S_0(t), t \geq 0$ for the case when the agent is not active, i.e. when $x(t) = 0$ for all t. This is the *unaffected* price process. When the strategy $x(t)$ is used, the price is changed from $S_0(t)$ to $S(t)$, and each market impact model has a particular way of describing this change. Typically, a pure buy strategy will lead to an increase of prices, and hence to $S(t) \geq S_0(t)$ for $t \in [0, T]$, while a pure sell strategy will decrease prices. The price impact is the main responsible for the *liquidation costs* [134] associated with the execution of metaorders. The other costs, i.e. commission fees and bid-ask spread costs, usually represent a small part of the liquidation costs regarding metaorders. These costs can be regarded as the difference between the cost that is paid during the execution and the mark-to-market value at the beginning of the execution $XS(0)$. For a continuous strategy this cost is given by

$$C[x(t)] = \int_0^T \dot{x}(t) (S(t) - S(0)) \, dt, \tag{6.4}$$

where $\dot{x}(s)$ is the rate of trading, i.e. number of shares per unit of time. This cost is also referred to as the *implementation shortfall* [202]. The problem of optimal order execution is to minimize costs in the class of all strategies that liquidate a given initial position of X shares during a predetermined time interval $[0, T]$. Optimality is usually understood in the sense that a certain risk functional is optimized, i.e. the cost functional reported in eq. 6.4 is only one of the possible choices. Commonly used risk functionals involve expected value of C as in Bertsimas & Lo (1998), Gatheral (2010) and others, mean-variance criteria as in Almgren & Chriss [15, 16], expected utility as in Schied [227], or alternative risk criteria as in Forsyth et al. [101] and Gatheral & Schied [105].

A minimal regularity condition is the requirement that the model does admit optimal order execution strategies. Moreover, the resulting strategies should be well-behaved. For instance, one would expect that an optimal execution strategy for a sell order should not involve intermediate buy orders and thus be a nonincreasing function of time. To make such regularity conditions independent of particular investors preferences, it is reasonable to formulate them in a risk-neutral manner, i.e., in terms of costs like reported in eq. 6.4. We will assume from now on that $S_0(t)$ is a martingale, when considering the regularity or irregularity of a market impact model. This condition is a standard assumption in the market impact literature, because

drift effects can be ignored [1] due to short trading horizons [16]. The first regularity condition was introduced by Huberman & Stanzl [143]. It concerns the absence of price manipulation strategies, which are defined as follows.

Definition 1 *A round trip is an order strategy $x(t)$ such that $\int_0^T \dot{x}(t)dt = 0$. A price manipulation strategy is a round trip with strictly negative expected cost*

$$\mathbb{E}\left[C\left[x(t)\right]\right] < 0. \tag{6.5}$$

A price manipulation strategy allows price impact to be exploited in a favorable manner. It was observed by Alfonsi et al. [14] that the absence of price manipulation may not be sufficient to guarantee the stability of a market impact model. There are models that do not admit price manipulation but for which optimal order execution strategies may oscillate strongly between buy and sell trades. This leads to the definition of *transaction-triggered price manipulation*.

Definition 2 *A market impact model admits transaction-triggered price manipulation if the expected costs of a sell (buy) program can be decreased by intermediate buy (sell) trades. That is, there exists $X \neq 0$, $T > 0$, and a corresponding order execution strategy $\hat{x}(t)$ for which*

$$\mathbb{E}\left[C\left[\hat{x}(t)\right]\right] < min_{x(t)\in M}\left\{\mathbb{E}\left[C\left[x(t)\right]\right]\right\}, \tag{6.6}$$

where M is the set of all monotone execution strategies $x(t)$ for X and T.

Yet another class of irregularities was introduced independently by Klock et al. [148] and Roch & Soner [216], i.e. *negative expected liquidation costs*.

Definition 3 *A market impact model admits negative expected liquidation costs if there exists $T > 0$ and a corresponding order strategy $x(t)$ for which*

$$\mathbb{E}\left[C\left[x(t)\right]\right] < 0. \tag{6.7}$$

The main difference with round trips is that in this case the volume X to execute can be also different from zero. The relations between the different notions of irregularities were given by Klock et al. [148].

Proposition 1 *(i) Suppose that asset prices are decreased by sell orders and increased by buy orders. The absence of transaction-triggered price manipulation implies that the model does not admit negative expected liquidation costs.*
(ii) Any market impact model that does not admit negative expected liquidation costs does also not admit price manipulation.

[1] If the trading proceeds are invested in an interest-bearing account, the drift term contributes to the return in excess of the risky asset. An alpha generating strategy should implement a price model with drift. The drift term is negligible if the execution of the metaorder occurs in a time period much smaller than the time period needed to generate the expected excess return, i.e. the asset's return minus the risk-free interest rate.

In particular, the absence of transaction-triggered price manipulation implies the absence of price manipulation in the Huberman & Stanzl sense. It is clear that the most restrictive condition is given by the absence of transaction-triggered price manipulations.

6.2.2 The Almgren-Chriss model

In this section we show the constraints implied by the absence of price manipulation for a specific market impact model, i.e. the Almgren-Chriss model [15, 16, 18]. It is one of the earliest market impact model classes that has so far been proposed, and which has also been widely used in the financial industry. In this model there are two impact components. The first component is *temporary* and only affects the individual trade that has also triggered it. The second component is *permanent* and affects all current and future trades equally. After the completion of the metaorder the price does not revert to zero. Order execution strategies $x(t)$, $t \in [0, T]$ are assumed to be absolutely continuous functions of time. Price impact of such strategies acts in an additive manner on a arithmetic Brownian unaffected asset prices. That is, for two nondecreasing functions $f, h : \mathbb{R} \to \mathbb{R}$, $g(0) = 0 = h(0)$

$$S(t) = S(0) + \int_0^t f(\dot{x}(s))\, ds + h(\dot{x}(t)) + \int_0^t \sigma\, dW(s), \tag{6.8}$$

where the term $h(\dot{x})$ corresponds to temporary, also called instantaneous, price impact, while the integral of $f(\dot{x})$ describes permanent price impact of single trades. Finally, σ is the volatility and $W(t)$ is a Wiener process. The term depending on $h(\dot{x})$ has no memory effect and only affects the current order. It is for instance caused by transaction costs or by crossing the bid-ask spread.

For the particular case $h(\dot{x}) = 0$, the next proposition was proved first by Huberman & Stanzl [143] in a discrete-time version of the Almgren-Chriss model and by Gatheral [104] in continuous time.

Proposition 2 *If an Almgren-Chriss model does not admit price manipulation for all $T > 0$, then f must be linear, i.e. $f(\dot{x}) = \alpha\dot{x}$ with a constant $\alpha \geq 0$.*

If $h(\dot{x}) \neq 0$, Gatheral & Schied [106] found the following result

Corollary 1 *Suppose that $f(\dot{x}) = \alpha\dot{x}$, for some $\alpha > 0$, and the function $\dot{x}h(\dot{x})$ is convex. Then the Almgren-Chriss model is free of transaction-triggered price manipulation, negative expected liquidation costs, and price manipulation.*

According to these constraints, the optimal strategy can be obtained by means of the classic calculus of variations. For linear functions $f(\dot{x}) = \alpha\dot{x}$ and $h(\dot{x}) = \beta\dot{x}$ the expected cost is given by

$$\mathbb{E}[C(x(t))] = \int_0^T \dot{x}(t)\,(\alpha\, x(t) + \beta\, \dot{x}(t))\, dt. \tag{6.9}$$

The optimal $x(t)$ solves the Euler-Lagrange equation

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{x}} - \frac{\partial L}{\partial x} = \ddot{x}(t) = 0, \tag{6.10}$$

where $L(x, \dot{x}) = \dot{x}(\alpha x + \beta \dot{x})$. With the boundaries conditions $x(0) = 0$, $x(T) = X$, the solution is given by a constant trading rate strategy $\dot{x}(t) = X/T$. This constant rate solution is known as VWAP solution[2].

The Almgren-Chriss model is highly tractable and can easily be generalized to multi-asset situations and to various risk criteria [16, 18, 171] and to nonlinear temporary market impact functions [18]. However, the conditions required by the absence of manipulations are quite strong. The main implication is a linear permanent impact function of trades, i.e. $f(\dot{x}) \propto \dot{x}$. This is in contrast with the empirical facts discussed in section 6.1, where a power law function is reported in eq. 6.1. The model showed in the next section is instead able to account for a power law market impact function.

6.2.3 The propagator model

In this section we show a model described by a *nonlinear transient* market impact. The transient impact is significant for a certain period after the placement of an order but vanishes eventually, due to the resilience effect of prices. It was proposed for the first time by Bouchaud et al. [43, 44] in order to conciliate a strong autocorrelation of the sign of trades [164] with a statistical efficient price, i.e. a diffusive price process. As showed by Lillo et al. [166], the autocorrelation of trade signs is a direct consequence of the slicing of metaorders. The paradox is that if the impact of each trade were permanent, the price process should be strongly super-diffusive due to the strong correlations among the signs of trades. The paradox can be solved introducing a price model where the impact of all trades is mediated by a decay function, i.e. the *propagator* function which gives the name to the model.

A continuous time version, together with an extensive analysis of the possible price manipulations, was developed by Gatheral [104]. This model is described by a price process of the form

$$S(t) = S(0) + \int_0^t f(\dot{x}(s))\, G(t-s)\, ds + \int_0^t \sigma\, dW(s), \tag{6.11}$$

where $\dot{x}(s)$ is the rate of trading, *i.e.* number of shares per unit of time, at time $s < t$, $f(\dot{x}(s))$ represents the impact of trading at time s, and the propagator $G(t-s)$ describes the impact decay. Finally, σ is the volatility and $W(t)$ is a Wiener process. Thus $S(t)$ follows an arithmetic random walk with a drift that depends on the accumulated impacts of previous trades. We refer to $f(\cdot)$ as the *instantaneous market impact function* and to $G(\cdot)$ as the *decay kernel*. An empirical analysis of this model is given in Lehalle & Dang [157]. This price model describes a metaorder impact

[2]VWAP stands for "volume-weighted average price ".

that asymptotically reverts to zero after the completion of the execution. The following results regarding the regularity of this model were given by Gatheral [104] and Gatheral et al. [107].

Proposition 3 *Suppose that* $G(t-s) = (t-s)^{-\gamma}$*, for some* $\gamma \in (0,1)$*, and that* $f(\dot{x}) = sign(\dot{x})\,|\dot{x}|^{\delta}$*, for some* $\delta > 0$*. Then price manipulation exists when* $\gamma + \delta < 1$ *and* $\gamma < \gamma^* = 2 - (\log 3/\log 2)$*.*

These parametric constraints are only a necessary condition, i.e. we do not know if price manipulation is possible for $\gamma + \delta > 1$ and $\gamma > \gamma^* = 2 - (\log 3/\log 2)$. Moreover, no-results are known about the presence or less of transaction-triggered price manipulations or negative costs in this region. One of the main purpose of the chapter 8 is to study what kind of irregularities can be present on this parametric region.

That it is necessary to consider decay kernels that are weakly singular [3], such as power-law decay, follows from the next result [107].

Proposition 4 *Suppose that* $G(\tau)$ *is finite and continuous at* $\tau = 0$ *and that* $f(\dot{x})$ *is nonlinear. Then the model admits price manipulation.*

A complete description of the optimal solution in the case of a linear transient market impact, i.e. where $f(\dot{x}) \propto \dot{x}$, was given by Gatheral et. al [108]. We discuss it in section 7.2.1. A description of the optimal solution in the case of a nonlinear impact function is still lacking. This was the motivation of the second part of this work, i.e. chapters 7, 8, in which we find optimal solution in presence of a nonlinear transient market impact. Moreover, we look for conditions that can possible extend the previous results regarding the regularity of optimal solutions for the model of eq. 6.11.

[3] A kernel is called singular if it becomes infinite at one or more points in the range of integration, such as in the Abel's equation [246]. A kernel is called weakly singular if its singularity is integrable, *i.e.* the integral of the function on a range that contains the singularity is finite. In our case the weakly singular kernel is given by $G(|t-s|) = |t-s|^{-\gamma}$.

Chapter 7

Optimal execution with nonlinear transient market impact: homotopy analysis approach

In this chapter we study the problem of the optimal execution of a large trade in the presence of nonlinear transient impact within the context of the propagator model [43]. The features of the price model and of the cost functional allow us to search for a statically optimal strategy that results also dinamically optimal. The optimal execution is defined by the solution of a nonlinear Urysohn integral equation of the first kind [207]. We propose an approach based on the homotopy analysis [139], whereby a well behaved initial strategy is continuously deformed to lower the expected execution cost. The deformations are described by the so-called deformation homotopy equations. We find that the optimal solution is front loaded for a concave impact and that its expected cost is significantly lower than that of conventional strategies.

7.1 Introduction

The optimization of trading strategies has long been an important goal for investors in financial markets. As demonstrated in the context of a linear equilibrium model by Kyle almost thirty years ago [153], the optimal strategy for an investor with insider information on the fundamental price of an asset is to trade incrementally through time. This strategy allows the trader to minimize costs whilst also minimizing the revelation of information to the rest of the market. The precise way in which it is optimal to split the large order (herein called *metaorder*) [254] depends on the objective function and on the market impact model, *i.e.* the change in price conditioned on signed trade size. In part due to the increasing tendency toward a full automation of exchanges and in part due to the discovery of new statistical regularities of the microstructure of financial markets, the problem of optimal execution is receiving growing attention from the academic and practitioner communities [10, 106].

As pointed out in Gatheral et al. [108], a first generation of market impact models [35, 15, 16, 18] distinguishes between two impact components. The first component

is temporary and only affects the individual trade that has triggered it. The second component is permanent and affects all current and future trades equally. These models can be either in discrete or in continuous time and can assume either linear or nonlinear market impact for individual trades. The second generation of market impact models focusses on the *transient* nature of market impact [43, 194, 44, 104]. In such models, market impact is typically assumed to factorize into two components: instantaneous market impact and a decay component. The instantaneous component models the reaction of price to traded volume. The decay component describes how the market price relaxes on average after the execution of an order. In such models, each trade affects future price dynamics with an intensity that decays with time.

The problem of optimal execution in the presence of transient impact has been considered in a series of recent studies. In the case of linear instantaneous market impact [108, 14, 54], the problem has been completely solved by showing that the cost minimization problem is equivalent to solving an integral equation. In particular Gatheral et al. [108] proved that optimal strategies can be characterized as measure-valued solutions of a Fredholm integral equation of the first kind. They show that optimal strategies always exist and are nonalternating between buy and sell trades when price impact decays as a convex function of time. This extends the result of Alfonsi et al. [14] on the non existence of transaction triggered price manipulation, *i.e* strategies where the expected execution costs of a sell (buy) program can be decreased by intermediate buy (sell) trades.

However, a series of empirical studies [165, 43, 26] has clearly shown that the instantaneous market impact is a strongly concave function of the volume, well approximated by a power law function. The resulting optimal execution problem in the presence of nonlinear and transient impact is mathematically much more complicated than the linear case.

Some important results in the nonlinear transient case were established by Gatheral [104] who showed that under certain conditions, the model admits price manipulation, *i.e.* the existence of round trip strategies with positive expected revenues. This money machine should of course be avoided in the modeling of market impact. In particular Gatheral set some necessary conditions for the absence of price manipulation (see below for details). A step toward the solution of the optimal execution problem under nonlinear transient impact has been made recently by Dang [81]. In his work, Dang suggests a way to convert the cost minimization problem into a nonlinear integral equation and proposes a numerical fixed point method on a discretization of the trading time interval to solve this equation. As we discuss in detail below, we find that Dang's fixed point method has convergence problems when the degree of nonlinearity of impact is significant and/or when the discretization grid is fine enough.

In this chapter we propose the homotopy analysis method (HAM) [160, 159, 162] to solve the optimal execution under nonlinear transient impact. The HAM is applied to the discretized version of the integral equation proposed by Dang [81]. The method starts from an initial guess and deforms it continuously in order to find better and better approximations of the solution of the integral equation. In doing this, we are implicitly restricting the space of solutions to continuous nonvanishing functions of

the trading rate. We find that the optimal solution is a non time-symmetric U-shape; in the case of concave (convex) instantaneous impact, it is optimal to trade more at the beginning (end) of the metaorder. A comparative cost analysis shows that our solution outperforms conventional strategies.

The homotopy method is now widely used in dealing with non-linear equations. Nonlinear equations are difficult to solve, especially analytically. Perturbation techniques are widely applied in science and engineering, and give a great contribution to help us understand nonlinear phenomena [192]. However, it is well known that perturbation methods are strongly dependent upon small/large physical parameters, and therefore are valid in principle only for weakly nonlinear problems. The so-called non-perturbation techniques, such as the Lyapunov's artificial small parameter method [175], the δ-expansion method [24], Adomian's decomposition method [11], and so on, are formally independent of small/large physical parameters. The problem is that all of these traditional non-perturbation methods can not ensure the convergence of solution series: they are in fact only valid for weakly nonlinear problems too. The homotopy analysis method [160] is a general analytic approach to get series solutions of various types of nonlinear equations, including algebraic equations, ordinary differential equations, partial differential equations, recently linear and nonlinear integral equations and coupled version of them. Unlike perturbation methods, the HAM is independent of small/large physical parameters, and thus is valid no matter whether a nonlinear problem contains small/large physical parameters or not. More importantly, differently from all perturbation and traditional non-perturbation methods, the HAM provides us a simple way to ensure the convergence of solution series, and therefore, the HAM is valid even for strongly nonlinear problems. More and more researchers have successfully applied this method to various nonlinear problems in science and engineering, such as the viscous flows of non-Newtonian fluids [244], the KdV-type equations [6], nonlinear heat transfer [7], finance problems [255, 200], projectile motion [251], magneto-hydrodynamics [13], Burgers-Huxley equation [188]. Abbasbandy with the co-authors [8] described the usage of HAM for solving the nonlinear Fredholm integral equation of the second kind. In this case the uniqueness of solution was proven and the sufficient condition for convergence of the created series was given.

In the case under analysis in this chapter we use the homotopy to study a numerical scheme to find an approximate solution to the integral equation.

The chapter is organized as follows. In Section 7.2, we state the problem and explain why it is difficult to solve. We also briefly summarize our results on the convergence of the Dang fixed point method. In Section 7.3, we present the HAM approach to the solution of the cost minimization problem.

7.2 The optimal execution problem and its solution

The model of nonlinear transient market impact for the price $S(t)$ of an order execution starting at time $t = 0$ when the price is $S(0) = S_0$ is

$$S(t) = S_0 + \int_0^t f(\dot{x}(s))\, G(t-s)\, ds + \int_0^t \sigma\, dW(s), \tag{7.1}$$

where $\dot{x}(s)$ is the rate of trading, *i.e.* number of shares per unit of time, at time $s < t$, $f(\dot{x}(s))$ represents the impact of trading at time s, and $G(t-s)$ describes the impact decay. Finally, σ is the volatility and $W(t)$ is a Wiener process. Thus $S(t)$ follows an arithmetic random walk with a drift that depends on the accumulated impacts of previous trades. We refer to $f(\cdot)$ as the *instantaneous market impact function* and to $G(\cdot)$ as the *decay kernel*. In discrete time this is the propagator model originally developed by Bouchaud et al. [44]; the above continuous time formulation (7.1) is due to Gatheral [104].

The optimal execution problem consists in finding the trading strategy $\Pi = \{x(t)\}_{t\in[0,T]}$ that minimizes the execution cost for a given total amount X of shares to be traded. The expected cost $C[\Pi]$ associated with a given strategy Π is given by

$$C[\Pi] = \mathbb{E}\left[\int_0^T \dot{x}(t)\,(S(t) - S_0)\, dt\right] = \int_0^T \dot{x}(t) \int_0^t f(\dot{x}(s))\, G(t-s)\, ds\, dt, \tag{7.2}$$

and the constraint that all shares should be traded is

$$\int_0^T \dot{x}(t)\ dt = X. \tag{7.3}$$

Expression (7.2) for the expected cost corresponds to expected implementation shortfall. We search for a statically optimal strategy. A statically optimal strategy is also dynamically optimal when the cost depends on the unaffected, i.e. without price impact, stock price only trough the term $\int_0^T S^u(t)\,\dot{x}(t)\ dt$, with $S^u(t)$ a martingale [210]. This implies that the cost function does not depend by the stochastic price motion. Thus for the model of (7.1), described by an unaffected martingale price process, and the cost function described by (7.2), a statically optimal strategy is also dynamically optimal.

The $\dot{x}(t)\, dt$ shares traded at time t are traded at an expected price

$$S(t) = S_0 + \int_0^t f(\dot{x}(s))\, G(t-s)\, ds, \tag{7.4}$$

which represents the cumulative impact of prior trading up to time t. It actually represents the impact of the metaorder if $t = T$, i.e. when computed on the entire trading period T.

The impact model of equation (7.1) is fully specified by the form of the functions f and G. A large body of empirical evidences points out two empirical facts on the form of these two functions. First, the instantaneous impact function $f(\cdot)$ is strongly concave. For example, based on a large sample of NYSE stocks, Lillo et al. [165] observed a concave function of the transaction volume. The concave function is well fitted by a power law with exponent 0.5 for small volumes and 0.2 for large volumes. Bouchaud et al. [44] analyzed stocks traded at the Paris Bourse and found that a logarithmic form gave the best fit to the data. In addition (in [44] for example), the decay kernel $G(\cdot)$ is found to decay asymptotically as a power law function

$$G(\tau) \sim \frac{1}{\tau^{\gamma}}. \tag{7.5}$$

The presence of these nonlinearities raises the question of the possible presence of price manipulation. There are different forms of manipulation. Following [106], we define a *price manipulation* as a round trip trade whose expected cost is negative. An impact model is free from price manipulation if, for any round trip trade, *i.e.* a strategy with $\int_0^T \dot{x}(t)\,dt = 0$, the expected cost is non negative. According to Proposition 1 of [107], the model of equation (7.1) admits price manipulation in the nonlinear case unless the decay kernel $G(\tau)$ is singular for $\tau = 0$. For these reasons we will focus our analysis on decay kernels of the form $G(t-s) = (t-s)^{-\gamma}$. Moreover, as shown by Gatheral [104], the requirement of no price manipulation restricts the class of possible joint form for the instantaneous impact and decay kernel. Specifically, for power law impact function, $f(\dot{x}) \propto sign(\dot{x})\,|\dot{x}|^{\delta}$, and a power-law kernel, $G(t-s) = (t-s)^{-\gamma}$, the following conditions

$$\gamma + \delta \geq 1, \qquad \gamma \geq \gamma^* = 2 - \frac{\log 3}{\log 2} \simeq 0.415, \tag{7.6}$$

are *necessary* for the absence of price manipulations. In this work we always consider parameters δ and γ satisfying the above conditions. However, if these conditions are satisfied, there is no guarantee that the impact model does not admit price manipulation. Later in the work we will show that it is indeed the case that the above conditions are not sufficient to preclude price manipulation. A weaker form of price manipulation, relevant for our work, is the *transaction-triggered* price manipulation defined in [14]. This occurs when the expected revenues of a buy(sell) program may be increased by intermediate sell(buy) trades.

7.2.1 The case of linear market impact

The optimization problem of minimizing the expected cost of equation (7.2) under the constraint of equation (7.3) in the case of linear impact, $f(\dot{x}) \propto \dot{x}$, has been solved and widely studied [108]. In what follows, we use the symbol $v(t)$ to indicate the rate of trading $\dot{x}(t)$. In particular, Proposition 22.9 of Ref. [108] states that if G is positive definite, then $x(t)$ minimizes the expected cost if and only if there is a λ

such that $\forall t$, $x(t)$ solves

$$\int_0^T G(|t-s|)dx(s) = \lambda \tag{7.7}$$

As an important example, relevant for this work, is the case $G(t-s) = (t-s)^{-\gamma}$ where the integral equation (7.7) becomes the Abel equation with solution

$$v(t) = \frac{c}{[t\ (T-t)]^{\frac{1-\gamma}{2}}}, \tag{7.8}$$

where c is uniquely determined by the constraint equation (7.3) as

$$c = X/\left(\sqrt{\pi}\left(\frac{T}{2}\right)^{\gamma}\frac{\Gamma\left(\left(1+\gamma\right)/2\right)}{\Gamma\left(1+\gamma/2\right)}\right), \tag{7.9}$$

where $\Gamma(\cdot)$ is Euler's Gamma function. This solution is U-shaped and symmetric under time reversal, *i.e.* $v(t) = v(T-t)$, $t \in [0, T/2]$. In the following we will refer to this solution as the GSS solution.

7.2.2 The general case of nonlinear market impact

In the general nonlinear case, the problem is mathematically much more complicated. A first step in this direction has been presented very recently by Dang [81] and consists in two contributions.

The first one is the use of calculus of variations for integrals depending on convolution products [223, 225, 224] to transform the cost minimization problem into an integral equation generalizing (7.7). Specifically, given $f \in C^1(\mathbb{R})$ and $G \in L^1[0,T]$, for the class of functions x on $[0,T]$ satisfying

- x is absolutely continuous on $(0,T)$,
- $f \circ v \in L^1[0,T]$,

the following necessary condition for the stationarity of the functional of equation (7.2) holds:

$$\int_0^t f(v(s))\, G(t-s)\, ds + f'(v(t)) \int_t^T v(s)\, G(s-t)\, ds = \lambda, \tag{7.10}$$

where again λ is a constant set by the constraint equation (7.3).

In the case of a convex impact function, $f(v) \propto sign(v)\,|v|^{\delta}$ with $\delta > 1$, equation (7.10) holds $\forall v \in \mathbb{R}$. In contrast, in the concave case, $\delta < 1$, equation (7.10) is not defined if the trading velocity v vanishes at some time t, because the derivative of f diverges at zero. This observation restricts the class of trajectories that can be considered. Moreover in the concave case there is no guarantee that the necessary condition (7.10) is also sufficient, because the minimization cost problem could have a large number of extremal points

Equation (7.10) is a weakly singular[1] Urysohn equation of the first kind [207] taking the form

$$\int_0^T G\left(|t-s|\right)\, F\left(v\left(s\right),t\right)\, ds = \lambda \tag{7.11}$$

where

$$F\left(v\left(s\right),t\right) = \begin{cases} f\left(v\left(s\right)\right), & s \leq t \\ v\left(s\right) f'\left(v\left(t\right)\right), & s > t. \end{cases} \tag{7.12}$$

Note that there are two sources of nonlinearity in the integral equation (7.11): the nonlinear impact function $f\left(v\right)$, and the function F. In fact, the nonlinearity depends also on the first derivative of the impact function f', *i.e.* on the response of price to the traded volume per unit time. Moreover, the term involving $f'(v(t))$ entangles the response of price at time t with the future trading rates, *i.e.* $v\left(s\right)$ for $s > t$, *i.e.* a coupling between present and future values of v. This means that equation (7.11) cannot be classified as a weakly singular nonlinear Fredholm equation, because the function F depends both on t and on s. In the linear impact case, both nonlinearities disappear and one recovers a weakly singular linear Fredholm integral equation of the first kind for the trading rate, where there is no such coupling between present and future times.

It is important to note that F in equation (7.11) is not an invertible function of v, because it depends not only on s but also on t. For this reason, the usual method for solving nonlinear integral equation by setting $u(t) = F(v(t))$ and solving the linear integral equation for u is not applicable here. In Section 7.3, we use the Homotopy Analysis Method to solve the integral equation (7.11).

Dang's fixed point algorithm

The second contribution of Dang [81] is a numerical scheme to solve the integral equation (7.10). This is a fixed point iteration scheme to find a numerical solution by quadrature methods which we review in Appendix 7.4.1. Quadrature relies on a discretization of the time interval $[0, T]$ into N subintervals, transforming the integral equation into a system of equations for a vector of dimension N. A detailed numerical analysis of the convergence properties of Dang's fixed point method as a function of the number of subintervals N and the deviation of the impact from linearity, as measured by $\delta - 1$ is presented in Appendix 7.4.1. We now summarize the results of this analysis.

The green area described in the left panel of Figure 7.1 shows the region where our implementation of Dang's algorithm converges, while the right panel shows the squared residual error of solutions (defined in equation (7.28) for a generic nonlinear operator) obtained as fixed points of the map. Both panels tell the same story. For small values of the number N of subintervals, the method converges, even if the

[1] A kernel is called singular if it becomes infinite at one or more points in the range of integration, such as in the Abel's equation [246]. A kernel is called weakly singular if its singularity is integrable, *i.e.* the integral of the function on a range that contains the singularity is finite. In our case the weakly singular kernel is given by $G\left(|t-s|\right) = |t-s|^{-\gamma}$.

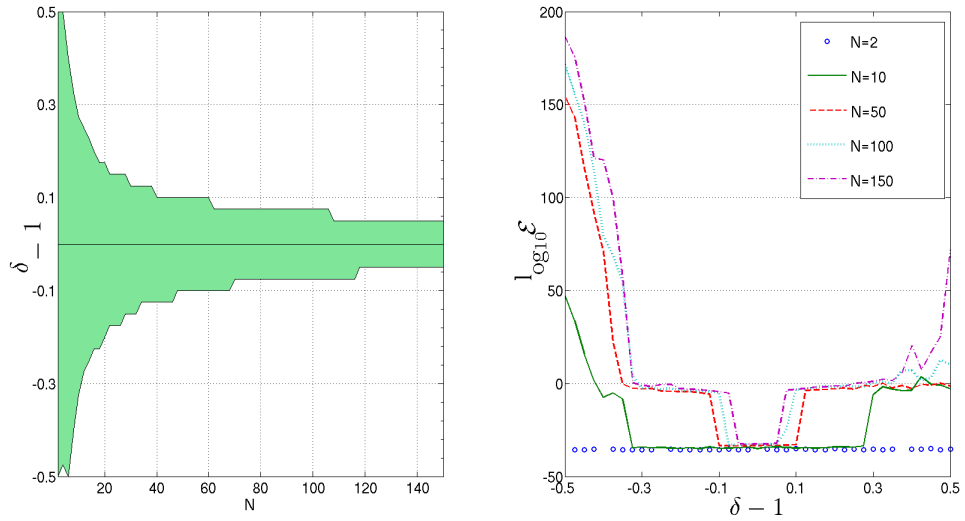

Figure 7.1: Left panel: convergence region of the Dang's fixed point method on the parameter space $(N, \delta - 1)$. Right panel: squared residual error of solutions obtained as map's fixed points.

impact function is strongly nonlinear. As N increases, *i.e* for finer partitions of the interval, we find that the method converges only for very moderately nonlinear impact functions. Dang's method obviously always converges when $\delta = 1$, because in this case, the objective function has only one minimum. That Dang's method does not converge is no surprise; iteration schemes are useful for solving integral equations of the second kind, whilst equation (7.11) is an integral equation of the first kind.

In summary, Dang's proposed solution technique in [81] is problematic. First, there is no guarantee that solving the integral equation (7.10) leads to a strategy that minimizes the expected execution cost (7.2) over the class of all admissible strategies. In particular , if the impact function is concave, as observed in reality, solutions are restricted to those for which the trading velocity $v(t)$ never vanishes, because $f'(v)$ diverges at $v = 0$. Second, according to our numerical tests, Dang's fixed point algorithm seems to converge only for moderate nonlinearity and moderate discretizations of the $[0, T]$ interval. Thus we need to investigate alternative approaches.

In the next Section we consider a perturbative approach to equation (7.10) valid for weak nonlinearity, *i.e.* $\delta \simeq 1$. This enables us to gain some understanding of the effect of nonlinearity on the optimal solution. In Section 7.3 we present a new approach to the solution of equation (7.10) using the homotopy method. Finally, in Section 8.2, we will present a brute-force numerical minimization of a discretized version of the cost function, where we will see explicitly that this cost function has many local minima.

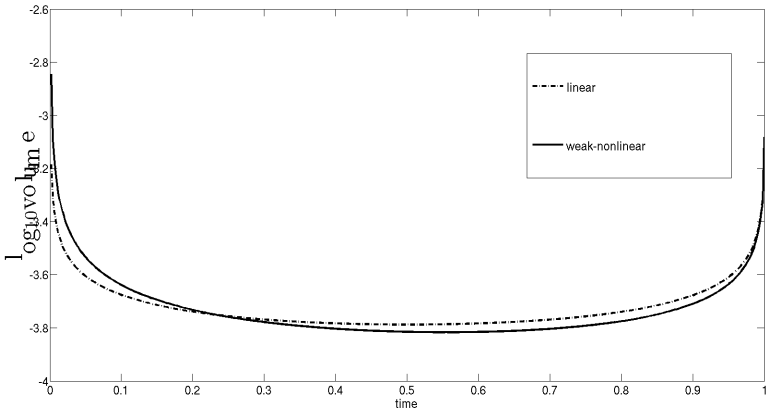

Figure 7.2: Solution of the Urysohn equation in the weak nonlinear case for $\gamma = 0.5, \epsilon = 0.02$ and $X = 0.1$. The full line represents the solution $v(s) = v_0(s) + \epsilon\, v_1(s)$. We observe that this solution is not symmetric under time reversal. The dotted line represents the GSS solution, *i.e.* the solution valid for the linear impact case.

7.2.3 A perturbative approach

In this section we present a simple perturbative method to investigate the solution of equation (7.10) in the weakly nonlinear case. This will provide some intuition for the results we obtain later using the more powerful homotopy method. Our perturbative approach is based on two approximations. The first one regards the impact function, while the second one regards the trading rate, *i.e.* the unknown solution.

Let us consider a buy program, $v(t) > 0$ and a slightly concave impact function, $f(v) = v^{1-\epsilon}$, with $0 < \epsilon \ll 1$. We then make the approximations

$$f(v) = v - \epsilon\, v\, \log(v) + \mathcal{O}\left(\epsilon^2\right); \qquad f'(v) = 1 - \epsilon - \epsilon\, \log(v) + \mathcal{O}\left(\epsilon^2\right). \tag{7.13}$$

Substituting into equation (7.10), and keeping terms only to order ϵ, we obtain

$$\begin{aligned} &\int_0^t v(s)\, G(t-s)\, ds + \int_t^T v(s)\, G(s-t)\, ds \\ &- \epsilon \left\{ \int_0^t v(s) \log(v(s))\, G(t-s)\, ds + [1 + \log(v(t))] \int_t^T v(s)\, G(s-t)\, ds \right\} = \lambda. \end{aligned} \tag{7.14}$$

As mentioned earlier, the GSS solution (7.8) solves the zeroth order case of (7.14)

$$\int_0^t v_0(s)\, G(t-s)\, ds + \int_t^T v_0(s)\, G(s-t)\, ds = \lambda_0, \tag{7.15}$$

in closed form.

Now write $v(s) = v_0(s) + \epsilon\, v_1(s) + \mathcal{O}(\epsilon^2)$. Then, matching terms of order ϵ gives the following linear Fredholm equation for $v_1(s)$:

$$\int_0^t v_1(s)\; G(t-s)\; ds + \int_t^T v_1(s)\; G(s-t)\; ds$$
$$= \int_0^t v_0(s)\; \log(v_0(s))\; G(t-s)\; ds + [1 + \log(v_0(t))] \int_t^T v_0(s)\; G(s-t)\; ds - \lambda'. \tag{7.16}$$

We solve the above equation for a fixed value of λ' using the constraint

$$\int_0^T [v_0(s) + \epsilon\, v_1(s)]\; ds = X$$

to determine the correct value of λ' for a fixed value of ϵ. We search for λ' iteratively until equation 7.16 is satisfied to within some given precision, in our case $1‰$ of X.

An example of this perturbative solution is shown in Figure 7.2. We use the quadrature method described in Section 7.3.2 to discretize the kernel G and perform matrix inversion to solve a discretized version of equation 7.16. We report here the case $\gamma = 0.5, \epsilon = 0.02$ and $X = 0.1$ using a quadrature grid with $N = 512$ points. Our first observation is that when market impact is concave, the optimal strategy is no longer symmetric under time reversal, in contrast to the linear case. Indeed, it is better to trade faster in the first half of the trading period T and more slowly in the second half. Repeating the computation in the convex market impact case, one obtains the opposite behavior: it is better to trade faster in the second half, as found also in [81] with the fixed point method. In Section 7.3, we will observe that the numerical solution of (7.10) in the strongly nonlinear impact case using the homotopy method has the same properties. It is worth emphasizing again that there is no reason to believe that a solution of the integral equation (7.10) gives the optimal execution strategy, corresponding to the global minimum of the expected cost (7.2).

7.3 The Homotopy Analysis Method

The concept of homotopy can be traced back to Henri Poincaré. In short, a homotopy describes a *continuous* variation or deformation. For example, a circle can be continuously deformed into a square or an ellipse, a coffee cup can be continuously deformed into a doughnut but not into a ball. More formally, let us consider the following general nonlinear equation:

$$\mathcal{N}[v(t)] = 0, \tag{7.17}$$

where $\mathcal{N}$ is a nonlinear operator, t denotes the independent variable, and $v(t)$ is the unknown function. Liao [160] constructs the so-called zero-order deformation equation

$$(1-p)\, \mathcal{L}\left[\phi(t;p) - v^0(t)\right] = p\, \hbar\, H(t)\, \mathcal{N}[\phi(t;p)], \tag{7.18}$$

where $p \in [0,1]$ is called the homotopy parameter, or embedding parameter and $\hbar$ is a non-zero auxiliary parameter which is called the convergence control parameter [9]. $H(t) \neq 0$ is an auxiliary function, $\mathcal{L}$ is an auxiliary linear operator, $v^0(t)$ is an initial guess of $v(t)$, and $\phi(t;p)$ is an unknown function. There are no particular prescriptions for the choice of auxiliary functions or operators; often the choice depends on the problem to be solved. When $p = 0$ and $p = 1$, we have respectively

$$\phi(t;0) = v^0(t), \qquad \phi(t;1) = v(t). \tag{7.19}$$

Thus, as p increases from 0 to 1, the solution $\phi(t;p)$ varies *continuously* from the initial guess $v^0(t)$ to the sought solution $v(t)$. Expanding $\phi(t;p)$ in a Maclaurin series with respect to p, we have

$$\phi(t;p) = v^0(t) + \sum_{m=1}^{\infty} v^m(t)\, p^m, \tag{7.20}$$

where

$$v^m(t) = \frac{1}{m!} \frac{\partial^m \phi(t;p)}{\partial p^m}\Big|_{p=0}. \tag{7.21}$$

The series representation of ϕ in equation (7.20) is called the homotopy series and $v^m(t)$ in (7.21) is called the mth-order homotopy derivative of ϕ [162]. If the auxiliary linear operator, the initial guess, the convergence control parameter $\hbar$, and the auxiliary function are properly chosen, the homotopy series converges at $p = 1$, giving the sought solution of equation (7.17). Then by using the relationship $\phi(t;1) = v(t)$, one has the so-called homotopy series solution

$$v(t) = v^0(t) + \sum_{m=1}^{\infty} v^m(t), \tag{7.22}$$

which is one of the solutions of the original nonlinear equation, as proved by Liao [161]. Defining the vector

$$\mathbf{v}^m = \left\{v^0(t), v^1(t), \cdots, v^m(t)\right\}, \tag{7.23}$$

and differentiating the zero-order deformation equation (7.18) m times with respect to the homotopy parameter p and then setting $p = 0$ and finally dividing them by $m!$, we have the so-called mth-order deformation equation

$$\mathcal{L}\left[v^m(t) - \chi^m v^{m-1}(t)\right] = \hbar\, H(t)\, R^m\left(\mathbf{v}^{m-1}\right), \tag{7.24}$$

where

$$R^m\left(\mathbf{v}^{m-1}\right) = \frac{1}{(m-1)!} \frac{\partial^{m-1} \mathcal{N}[\phi(t;p)]}{\partial p^{m-1}}\Big|_{p=0} \tag{7.25}$$

and

$$\chi^m = \begin{cases} 0, & m \leq 1\,, \\ 1, & m > 1. \end{cases} \tag{7.26}$$

The nth-order approximate solution is then given by

$$v^{(n)}(t) = v^0(t) + \sum_{m=1}^{n} v^m(t)\,, \tag{7.27}$$

and the exact solution by $v(t) = \lim_{n\to\infty} v^{(n)}(t)$. The two main difficulties of this approach are to compute the derivatives of equation (7.25) and how to choose an appropriate value of $\hbar$ in order to guarantee the convergence of the series solution of equation (7.22).

The first problem requires us to compute the homotopy derivative of a given nonlinear smooth function f of the homotopy Maclaurin series of equation (7.20), *i.e.* to compute the derivative of $f(\phi)$. We refer the reader to Appendix 7.4.2 for details. The second problem is how to choose an appropriate value of the convergence control parameter $\hbar$ in order to guarantee the convergence of the series of equation (7.22) [168]. To do this, we will adopt in the following, the so-called *optimization* method [162, 13], according to which, we define the squared residual of the governing equation (7.17) as

$$\mathcal{E}^n(\hbar) = \int_0^T \left(\mathcal{N}\left[v^{(n)}(t)\right]\right)^2 dt. \tag{7.28}$$

The optimal value of the convergence control parameter is then obtained by finding the minimum of this squared residual. In fact, if $v^{(n)}(t)$ is the solution of the original problem of equation (7.17), the residual $\mathcal{E}^n(\hbar)$ vanishes.

7.3.1 Homotopy for nonlinear transient market impact

Now we apply the Homotopy Analysis Method (HAM) to the solution of the nonlinear integral equation (7.11). Thus

$$\mathcal{N}[v(t)] = -\lambda + \int_0^T G(|t-s|)\; F(v(s), t)\; ds. \tag{7.29}$$

As suggested in [**?**], we choose the linear operator $\mathcal{L}$ to be the identity, that is

$$\mathcal{L}[\phi(t;p)] = \phi(t;p)\,, \tag{7.30}$$

and the auxiliary function to be $H(t) = 1$. The zero-order deformation equation is then

$$(1-p)\left[\phi(t;p) - v^0(t)\right] = \hbar\, p\, \mathcal{N}[\phi(t;p)]\,. \tag{7.31}$$

Differentiating this zero-order deformation equation m times with respect to p, and finally dividing by $m!$, we obtain the mth-order deformation equation

$$v^m(t) = \chi^m v^{m-1}(t) + \hbar R^m\left(\mathbf{v}^{m-1}\right), \tag{7.32}$$

where for $m > 1$

$$\begin{aligned} R^m\left(\mathbf{v}^{m-1}\right) &= \frac{1}{(m-1)!}\frac{\partial^{m-1}\mathcal{N}\left[\phi(t;p)\right]}{\partial p^{m-1}}\Big|_{p=0} \\ &= \int_0^T G(|t-s|)\left\{\frac{1}{(m-1)!}\left.\frac{\partial^{m-1}F\left(\phi(s;p),t\right)}{\partial p^{m-1}}\right|_{p=0}\right\} ds\,. \end{aligned} \tag{7.33}$$

For example, the first-order deformation equation is

$$v^1(t) = \hbar\left(-\lambda + \int_0^T G(|t-s|)\,F\left(v^0(s),t\right)\,ds\right). \tag{7.34}$$

To compute the higher-order deformation equations we need to write F as a function of two homotopy-series, defined respectively at the time points s and t

$$F\left(\sum_{i=0}^{\infty} v^i(s)\,p^i, \sum_{j=0}^{\infty} v^j(t)\,p^j\right) = \begin{cases} f\left(\sum_{i=0}^{\infty} v^i(s)\,p^i\right), & s \le t \\ \left(\sum_{i=0}^{\infty} v^i(s)\,p^i\right)\; f'\left(\sum_{j=0}^{\infty} v^j(t)\,p^j\right), & s > t. \end{cases} \tag{7.35}$$

and then apply the homotopy-derivative of equation (7.57), relative to a nonlinear one-dimensional system, for $s \le t$, and the homotopy-derivative of equation (7.58), relative to a nonlinear two-dimensional system, for $s > t$. The expression of the first homotopy-derivative, for $v^0(t) > 0$, is given by

$$\frac{\partial}{\partial p}F\left(\sum_{i=0}^{\infty} v^i(s)\,p^i, \sum_{j=0}^{\infty} v^j(t)\,p^j\right)\Bigg|_{p=0} = \begin{cases} \delta\left(v^0(s)\right)^{\delta-1} v^1(s), & s \le t \\ \delta\, v^1(s)\left(v^0(t)\right)^{\delta-1} + \delta(\delta-1)\,v^0(s)\left(v^0(t)\right)^{\delta-2} v^1(t), & s > t. \end{cases} \tag{7.36}$$

Higher orders may be computed using symbolic computation software such as Mathematica or Maple.

Obviously, for this algorithm to work, the initial guess needs to satisfy the condition $f'(v^0(t)) < \infty$ for all $t \in [0,T]$. Thus, v^0 must have the same sign on the whole interval $[0,T]$. From now on, we consider a buy program. In order to test for possible dependencies on the initial guess, we choose both the VWAP strategy, $v^0_{VWAP}(t) = X/T$, and the GSS solution for the linear case, *i.e.* v^0_{GSS} given by equation (7.8) as initial guesses.

7.3.2 A Discrete Homotopy Analysis Method

To apply HAM to our problem, we need to compute the definite integrals (7.33), which seem to be analytically intractable. We therefore propose a way to approximate these integrals, and refer to this discretized version of HAM as the DHAM approach.

We first discretize equation (7.11) by splitting the time interval $[0, T]$ into N subintervals at the times $t_i = i\,T/N$ where $i \in \{0, 1, \cdots, N\}$. This gives the following nonlinear system of N equations in the variables $v_i = v\,(t_i)$ where $i \in \{1, \cdots, N\}$

$$\sum_{j=1}^{N} G_{ij} F_{ij}\,(v) = \lambda \tag{7.37}$$

where i indicates the time point t_i. The nonlinear function $F\,(.)$ of equation (7.12) becomes a real $N \times N$ matrix

$$F_{ij}\,(v) = \begin{cases} f\,(v_j)\,, & j \leq i \\ v_j f'\,(v_i)\,, & j > i. \end{cases} \tag{7.38}$$

The decay kernel $G\,(|t-s|)$ becomes a Toeplitz real symmetric $N \times N$ matrix given by

$$G_{ij} = \int_{t_{i-1}}^{t_i} \int_{t_{j-1}}^{t_j} G\,(|t-s|)\; ds\; dt. \tag{7.39}$$

If $G\,(\tau) = \tau^{-\gamma}$, for $i > j$ we have

$$G_{ij} = \frac{1}{(1-\gamma)\,(2-\gamma)} \left(\frac{T}{N}\right)^{2-\gamma} \left\{(i-j+1)^{2-\gamma} - 2\,(i-j)^{2-\gamma} + (i-j-1)^{2-\gamma}\right\} \tag{7.40}$$

and the diagonal terms are given by

$$G_{ii} = \frac{2}{(1-\gamma)\,(2-\gamma)} \left(\frac{T}{N}\right)^{2-\gamma}. \tag{7.41}$$

In this scheme the constraint on the traded volume of equation (7.3) is given by

$$\sum_{i=1}^{N} v_i = \frac{NX}{T}. \tag{7.42}$$

We are not able to solve the nonlinear system of equation (7.37) directly. Rather, we search an approximate homotopy series solution

$$v_i = v_i^0 + \sum_{m=1}^{\infty} v_i^m, \;\; i \in \{1, \cdots, N\}\,, \tag{7.43}$$

where the deformation equations (7.32) for $m > 1$ are approximated by

$$v_i^m = v_i^{m-1} + \hbar \sum_{j=1}^{N} G_{ij}\, F_{ij}^{m-1}, \tag{7.44}$$

and where the $(m-1)$-th homotopy derivative is evaluated on the grid points t_i, s_j

$$F_{ij}^{m-1} = \frac{\partial^{m-1}}{\partial p^{m-1}} F\left(\sum_{k=0}^{\infty} v_i^k\, p^k, \sum_{l=0}^{\infty} v_j^l\, p^l \right) \Big|_{p=0}. \tag{7.45}$$

The approximate solution of order n is given by

$$v_i^{(n)} = v_i^0 + \sum_{m=1}^{n} v_i^m, \tag{7.46}$$

and we compute the squared residual error of equation (7.37) as

$$\mathcal{E}^n(\hbar) = \sum_{i=1}^{N} \left[-\lambda + \sum_{j=1}^{N} G_{ij}\, F_{ij}\left(v^{(n)}\right) \right]^2. \tag{7.47}$$

The value of $\hbar_{min}$ that minimizes this error gives the DHAM solution $v_i^{(n)}(\hbar_{min})$ of equation (7.11). This solution can be considered as a piecewise constant approximation of the exact solution corresponding to a sequence of VWAP executions with trading rates $v_i^{(n)}$.

Finally, the expected liquidation cost (7.2) is approximated by

$$C\left[v^{(n)}\right] = \sum_{i=1}^{N} \sum_{j=1}^{N} v_i^{(n)}\, f\left(v_j^{(n)}\right) A_{ij}, \tag{7.48}$$

where the A_{ij} are elements of a Toeplitz matrix that describes the decay kernel $G(t-s)$

$$\begin{aligned} A_{ij} &= 0;\ j > i, \\ A_{ii} &= G_{ii}/2; \\ A_{ij} &= G_{ij};\ j \leq i. \end{aligned} \tag{7.49}$$

7.3.3 DHAM results

We present here the optimal strategies obtained with DHAM in the no-dynamic-arbitrage region given by equation (7.6). We analyze in detail the values $\gamma = 0.45$ and $\gamma = 0.5$ in a strong nonlinear regime, i.e. $\delta = 0.5$. We consider the case where the volume to be purchased is $X = 0.1$, which can be interpreted as a metaorder execution where one buys 10% of the available unitary market volume. We discretize

using a grid of $N = 100$ subintervals and compute the solution up to the 7-th order with a tolerance on the constraint on the total quantity executed of 1‰ of X. As mentioned earlier, we consider two initial guesses, namely one corresponding to a VWAP profile and the other to a GSS solution (7.8).

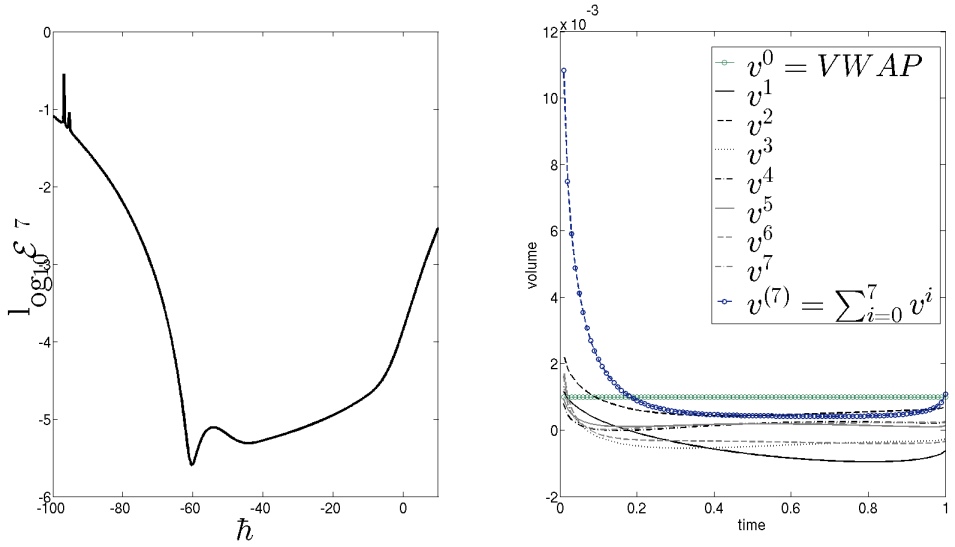

Figure 7.3: The logarithm of the squared residual $\mathcal{E}^7(\hbar)$ is illustrated on the left panel, the minimum is attained for $\hbar = -60.3$ where we have $\mathcal{E}^7 = 2.5 \times 10^{-6}$. The VWAP initial guess and the DHAM solution are reported on the right panel respectively by a full green line with circles and a dashed blue line with circles, are reported also the results of the seven deformation equations.

The results for a VWAP initial condition are illustrated in Figure 7.3. The left panel shows that the minimum of squared residual of the 7-th order iteration is attained at $\hbar = -60.3$. The results for a GSS initial condition are illustrated in Figure 7.4, where the minimum of the squared residual is attained at $\hbar = -55.7$ (left panel). The right panel of both figures shows the trading profile of the different approximating terms v^i, $(i = 1, ..., 7)$, as well as of the 7-th order approximated solution $v^{(7)}$ obtained by summing the terms. It is worth noticing that higher order terms become smaller and smaller. In both cases the DHAM approximated solution is asymmetric with respect to time reversal; it is optimal to trade faster at the beginning of the trading period and slower at the end. This is consistent with our earlier findings in Section 7.2.3 for small deviations from linear impact. Finally, it is worth noticing that the two different initial conditions lead to very similar approximated solutions. In fact, if we exclude the last discretization point ($t_{100} = T$), the maximum value of the relative difference between the two profiles is around 8%, but the relative difference of the expected liquidation cost, given by equation (7.48), is only 0.5%.

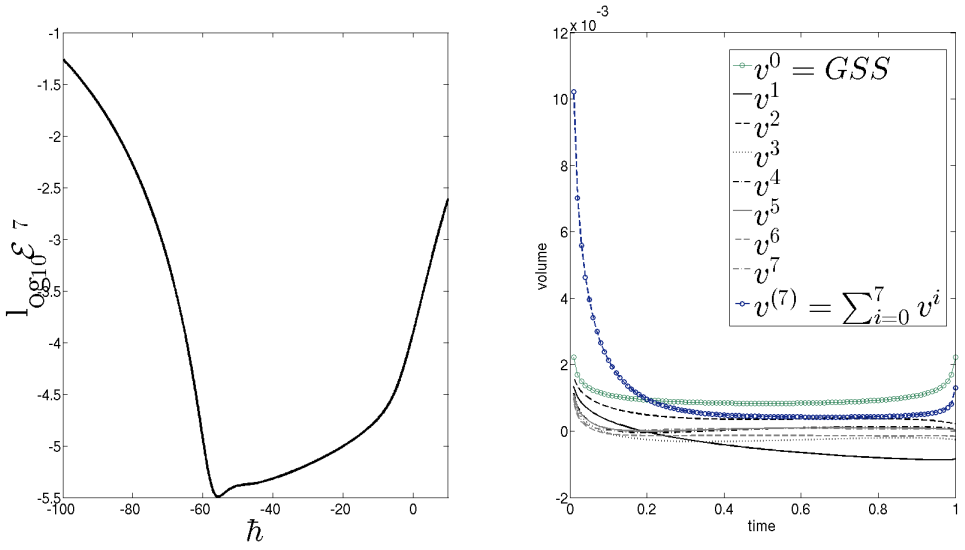

Figure 7.4: The logarithm of the squared residual $\mathcal{E}^7(\hbar)$ is illustrated on the left panel, the minimum is attained for $\hbar = -55.7$ where we have $\mathcal{E}^7 = 3.2 \times 10^{-6}$. The GSS initial guess and the DHAM solution are reported on the right panel respectively by a full green line with circles and a dashed blue line with circles, are reported also the results of the seven deformation equations.

We now discuss the advantage of a DHAM execution with respect to VWAP[2] and GSS executions in terms of expected liquidation costs. The expected liquidation costs of the three strategies are reported in Table 7.1, where we consider $\gamma = 0.45$ and 0.5 and $\delta \in [1/2, 1]$. As expected, close to the linear case ($\delta = 1$) the difference in cost between the DHAM and the GSS is negligible[3]. Interestingly, the VWAP strategy, which is clearly different from the GSS, has a cost essentially equivalent. Thus in the linear case, the advantage of sophisticated execution strategies relative to a straightforward VWAP is very small. In contrast, as we move toward the strongly nonlinear regime, *i.e.* $\delta \approx 0.5$, the DHAM solution achieves the smallest cost, while the worst strategy is the VWAP. As can be observed from Table 7.1, the improvement of DHAM with respect to GSS is much larger than the improvement of GSS with respect to VWAP. For example, for $\delta = \gamma = 0.5$, the DHAM has a cost 20% smaller than the GSS, while the latter has a cost which is only 1% smaller than the VWAP.

To summarize, solutions using the DHAM approach display the same time asymmetry that we found earlier using a simple perturbative approach. More importantly,

[2]The cost of a VWAP execution is $C_{VWAP} = X\, f(X)\, T^{(1-\gamma)} / ((1-\gamma)(2-\gamma))$.

[3]In the linear case, i.e. $\delta = 1$, is possible to find a DHAM strategy with a cost lower than that of GSS strategy. This is only a numerical artifact, because here we implement a discretized version of the GSS strategy defined by (7.8). We compute the cost of a constant piece-wise approximation of the GSS. This approximation is then used as the initial guess of the DHAM solution. In the linear case, when the value of N increases the costs of the approximated GSS and of DHAM strategies become equal.

	VWAP	GSS	DHAM	VWAP	GSS	DHAM
δ	$\gamma = 0.45$	$\gamma = 0.45$	$\gamma = 0.45$	$\gamma = 0.5$	$\gamma = 0.5$	$\gamma = 0.5$
1.0	0.0117	**0.0116**	**0.0116**	0.0133	0.0132	**0.0131**
0.95	0.0132	**0.0130**	**0.0130**	0.0150	**0.0148**	**0.0148**
0.90	0.0148	0.0146	**0.0143**	0.0168	0.0166	**0.0164**
0.85	0.0166	0.0164	**0.0162**	0.0188	0.0186	**0.0185**
0.80	0.0186	0.0184	**0.0179**	0.0211	0.0209	**0.0204**
0.75	0.0209	0.0206	**0.0198**	0.0237	0.0234	**0.0227**
0.70	0.0234	0.0231	**0.0218**	0.0266	0.0263	**0.0249**
0.65	0.0263	0.0260	**0.0235**	0.0298	0.0295	**0.0274**
0.60	0.0295	0.0291	**0.0251**	0.0335	0.0331	**0.0297**
0.55	0.0331	0.0327	**0.0275**	0.0376	0.0372	**0.0323**
0.50				0.0422	0.0417	**0.0347**

Table 7.1: Costs for three different strategies, VWAP, GSS, and DHAM, in the no-dynamic-arbitrage region for $\gamma = 0.45, 0.5$. The numbers in boldface indicate strategies achieving the lowest expected cost. The difference between expected costs increases with the degree of non-linearity. In each case we use a GSS initial guess to obtain the DHAM solution.

the DHAM approach allows us to compute the optimal strategy in the strong non-linear regime, achieving expected execution costs which are significantly smaller than those obtained with VWAP or GSS solutions. Once again however, the DHAM solution does not necessarily correspond to a global minimum of the cost functional (7.2), because it is a continuous deformation of a continuous and positive initial guess. In order to find potentially lower cost strategies, in Section 8.2, we will tackle the problem of direct numerical minimization of the cost function.

7.4 Appendices

7.4.1 Dang's fixed point algorithm

In this section we analyze an iteration scheme proposed by Dang [81] to find a numerical solution of equation (7.37) by quadrature methods. Dang considers a power law impact function $f(v) = sign(v)|v|^{\delta}$ where $\delta > 1$. If we want to study the case $\delta < 1$ we should find a way to handle the problem of an infinite first derivative of $f'(v)$ for $v = 0$. We define a perturbative version of the market impact function $\hat{f}(v)$

$$\begin{aligned} \hat{f}(v) &= sign(v)(\epsilon + |v|)^{\delta} \\ \hat{f}'(v) &= \delta(\epsilon + |v|)^{\delta-1} \end{aligned} \tag{7.50}$$

where the perturbation is given by the parameter $0 < \epsilon \ll 1$. In this way the trading rate can assume any value on $\mathbb{R}$ We search a solution $v_i(\epsilon)$ of the equation (7.37) with a perturbation near to zero, *i.e.* $\epsilon \approx 0$.

Dang defines a nonlinear map from equation (7.37) and searches a fixed point for the map[7] starting from the initial guess v_i^0. This means that we have a sequence of approximations for the solution $v_i^0, v_i^1, v_i^2, \cdots$, the convergence to a fixed point of the map is a necessary condition in order to have a convergent iteration scheme [33, 51]. The iterative scheme is defined by a Taylor expansion

$$F_{ij}\left(v^m\right) \approx F_{ij}\left(v^{m-1}\right) + \left(v_j^m - v_j^{m-1}\right) F'_{ij}\left(v^{m-1}\right) , \tag{7.51}$$

where

$$F'_{ij}\left(v\right) = \begin{cases} \hat{f}'\left(v_j\right), & \text{if } j \leq i \\ \hat{f}'\left(v_i\right), & \text{if } j > i \end{cases} \tag{7.52}$$

is the derivative[8] of $F_{ij}(v)$ in equation (7.38) with respect to v_j. Following Dang, we solve for each m-th iteration the following linear system:

$$\boldsymbol{K} v^m = \tilde{c}. \tag{7.53}$$

This defines a nonlinear map M

$$v^m = \boldsymbol{K}^{-1}\left(v^{m-1}\right)\tilde{c}\left(v^{m-1}\right) = M\left(v^{m-1}\right) , \tag{7.54}$$

where

$$\begin{aligned} K_{ij} &= G_{ij} F'_{ij}\left(v^{m-1}\right) , \\ \tilde{c}_i &= \lambda - \sum_{j=1}^{i} G_{ij}\left(F_{ij}\left(v^{m-1}\right) - v_j^{m-1} F'_{ij}\left(v^{m-1}\right)\right) . \end{aligned} \tag{7.55}$$

We search a solution that is a fixed point of the map M, *i.e.* $v^* = M\left(v^*\right)$, as in Dang [81]. Starting from the initial guess v^0 we look for a value $\bar{m}$ for which the N-dimensional vector v^m is a constant vector for $m > \bar{m}$. A simple scalar quantity controlling if the fixed point has been reached is the mean field, defined as $\bar{v}^m = 1/N \sum_{i=1}^{N} v_i^m$. The mean field $\bar{v}$ is also useful in order to implement a procedure that leads to determine the value of λ that satisfies the constraint (7.42) on the total quantity traded. It is clear that for a general dynamic system the condition of a constant mean field $\bar{v}$ does not guarantee the existence of a fixed point since there can be also more complex attractors, *e.g.* a chaotic attractor, where the mean field is constant. Here we are interested in studying only a necessary condition in order to find a fixed point of the nonlinear map M. Moreover, it is important to notice that the evolution of the map M does not satisfy the constraint of equation (7.42), *i.e.* if

[7] We have used different discretization methods to approximate the integral equation, including the product Nystrom method [**?**] used by Dang. We find that our conclusions do not depend on the kind of discretization used.

[8] Here we can observe the fundamental difference between our DHAM approach and Dang's fixed point approach. The DHAM approach is based on the observation that this dynamical system has to be regarded as a bi-dimensional system; in contrast, Dang develops a method that lacks interaction between past and present values of the trading rate v.

the initial point v^0 is on the hyper-plane, it could go outside of it when we iterate the map M. This is the reason for which we should tune the constant λ in order to satisfy our constraint within some precision.

Dang's proposal for the initial guess $v_i^0 = v$, $\forall i$, is a constant vector whose components are given by

$$v f'(v) \sum_{j=1}^{N} G_{1j} = \lambda. \tag{7.56}$$

In principle the fixed point could depend on the starting point v^0 that we choose as initial guess for the solution. We add to this initial guess a set of initial guesses uniformly distributed on the simplex to investigate the presence of different attractors. The convergence of this method is determined by two parameters: N and δ.

We performed an extensive analysis of the convergence of the Dang's fixed point method. The convergence criterion is that the relative standard deviation of the mean field is below some threshold, typically 10^{-9}. The result of this analysis for $\gamma = 0.5$ and $\epsilon = 0$ is shown in Figure 7.1, but similar results are obtained for small values of ϵ. We observe that when N increases, the set of values of δ for which we have the convergence decreases, *i.e.* we have a δ_{min} under which the procedure does not converge, *i.e.* we do not find a fixed point for the map M, see Figure 7.1. When the procedure converges, the solution does not depend from the initial guess, i.e the map M has only one attractive basin. Moreover, the convergence is extremely fast, reaching the fixed point in very few iterations of the map. As N increases however, Dang's algorithm converges only in the weakly nonlinear case.

Confining our attention to a regime where Dang's algorithm does converge, how does the solution achieved with Dang's algorithm compare with the solution achieved using the DHAM and SQP methods we propose? In Figure 7.5 we report results for the very weakly nonlinear case with $N = 100, T = 1, X = 0.1, \gamma = 0.5, \delta = 0.95, \epsilon = 10^{-6}$. We find that in this case the Dang method converges very quickly (right panel) and the solution is essentially coincident with the one obtained with the DHAM method (left panel). The solution of the numerical optimization using the SQP algorithm is oscillating, alternating between positive and negative trading rates.

In conclusion, Dang's algorithm converges only under weak nonlinearity and/or when discretization of the time interval is very coarse. When the Dang method does converge, the solution is very close to the one obtained with the DHAM method proposed in this work. However the DHAM approach is applicable not only in the weakly nonlinear regime where Dang's algorithm converges, but also in the strongly nonlinear regime.

7.4.2 Homotopy derivatives

As explained in Section 7.3, the so-called homotopy derivative is used to deduce deformation equations of equation (7.24) for any order greater than one, *i.e.* to compute v^m for $m > 1$. Such computations are difficult because they depend on the nonlinear operator $\mathcal{N}$. For our problem, we need to compute the homotopy derivative of a power law function.

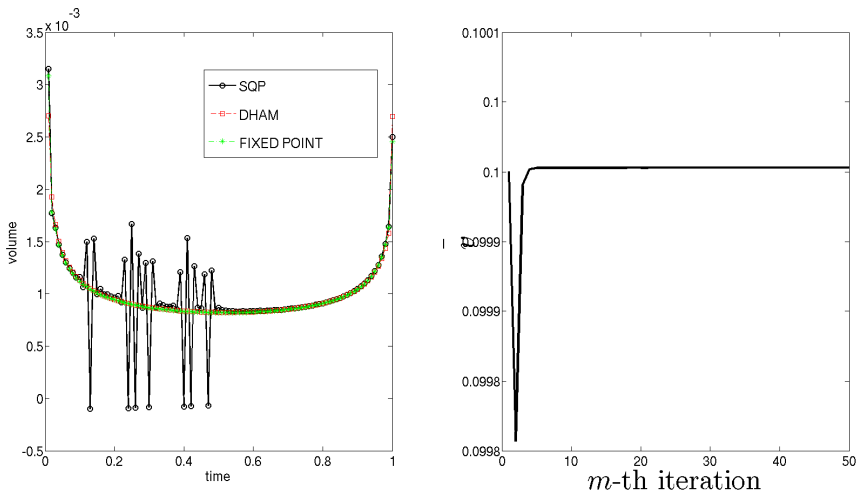

Figure 7.5: Left panel. Optimal solutions using the Dang fixed point method, the DHAM and the SQP method in the weakly nonlinear regime: $\gamma = 0.5, \delta = 0.95, N = 100, T = 1, X = 0.1$. The right panel reports the fast convergence of the mean field $\bar{v}^m$ of the fixed point method with a VWAP initial guess and $\lambda = 2.87 * 10^{-3}$. Notice that the initial guess goes outside the constraint at the beginning of the iteration procedure.

A nonlinear power law function with a integer exponent, *i.e.* $f(\phi) = \phi^k,\ k \in \mathbb{N}$, was studied by Molabahrami and Khani [188] to find an approximate solution of the Burgers-Huxley equation. The case of a real power-law index was studied by Wang et al. [244] analyzing the flow of a power-law fluid film on an unsteady stretching surface. Our case (7.12) is more complicated than the case of a simple power-law function. Recent results given by Turkyilmazoglu [240] and Liao [162] show how to compute the homotopy-derivative of any smooth function. The homotopy-derivative $\mathcal{D}^m[f(\phi)] = (1/m!)\,\partial^m f(\phi)/\partial p^m$ is given by the recursive relation

$$\mathcal{D}^m[f(\phi)] = \sum_{k=0}^{m-1}\left(1 - \frac{k}{m}\right)\mathcal{D}^{m-k}[\phi]\,\frac{\partial}{\partial\phi}\left\{\mathcal{D}^k[f(\phi)]\right\}, \tag{7.57}$$

evaluated at $p = 0$. The sum consists of two homotopy-derivative terms. The first gives v^{m-k}, the second term gives polynomial terms of v^i that multiply $f^i(v^0),\ i = 1, \cdots, m$, derivatives of f evaluated at the initial guess. If the market impact function f is of the form $f(v) \propto v^\delta$ with $\delta < 1$, all such derivatives diverge at $v = 0$. As mentioned in Section 7.3, we avoid this problem by choosing the initial guess to be a strictly positive (or negative) function of time. Using the $m-1$-th order of equation (7.57) in equation (7.25) we can express the m-th homotopy derivative v^m as a complicated function of the previous $m-1$ derivatives.

To handle the nonlinearity of equation (7.12) in the HAM framework we need a further step. The coupling between past and future values of trading rates implies that we have to consider our problem as a two-dimensional system. This means that we use the homotopy-derivative for systems described by two variables [162], for example u and w, in which we have a nonlinear coupling between them given by $f(u, w)$

$$\mathcal{D}^m \left[f(\phi, \psi) \right] = \sum_{k=0}^{m-1} \left(1 - \frac{k}{m} \right) u^{m-k} \mathcal{D}^k \left[\frac{\partial f(\phi, \psi)}{\partial \phi} \right] + \left(1 - \frac{k}{m} \right) w^{m-k} \mathcal{D}^k \left[\frac{\partial f(\phi, \psi)}{\partial \psi} \right], \tag{7.58}$$

where ϕ, ψ are the Maclaurin series of u, w respectively. Thus, given the expression of equation (7.12) as $F(v(s), v(t))$, we are able to apply the HAM to our optimal execution problem using the equation (7.58) and choosing a function with a given sign as initial guess.

Chapter 8

Optimal execution with nonlinear transient market impact: numerical optimization approach

In this chapter we consider a brute force numerical optimization of the cost functional of eq. 7.48 relative to the propagator model [43, 104]. The purpose of this analysis is to search for strategies not described by the Urysohn equation 7.11. Optimal, or sub-optimal, strategies are defined as local minimizers of the cost functional. The cost landscape appear to be described by many minima. We find that the optimal solution for a buy program typically features a few short intense buying periods separated by long periods of weak selling. Such strategies have expected costs lower than that described tby a pure buy strategy, i.e. we find evidence of transaction triggered price manipulations. In some cases, i.e high non-linearity, we find also the presence of negative expected cost. We show that this undesirable characteristics of the propagator model may be mitigated either by introducing a bid-ask spread cost or by imposing convexity of the instantaneous market impact function for large trading rates.

8.1 Introduction

The HAM method explores only a restricted subspace of possible solutions of the optimization problem described in section 7.2. Such solutions are defined by the stationarity condition given by the Urysohn equation 7.11. For this reason in this chapter we consider a fully numerical cost optimization method on a discrete grid. By using Sequential Quadratic Programming (SQP) we minimize directly the cost functional (*i.e.* we do not try to solve the integral equation). We find that the cost landscape is rugged, *i.e.* composed by a very large number of local minima separated by peaks. A significant number of these minima correspond to strategies with similar costs; for a buy program the corresponding strategy is an alternation of intense and short bursts of buying periods and long periods of weak selling. In other words, the model admits transaction triggered price manipulation. More important, when

the nonlinearity is strong and/or the partition is fine, some such strategies have a negative expected cost, indicating that the model admits price manipulation. We then further extend our analysis by minimizing the cost functional with the extra constraint that all trades should have the same sign, so that for example selling is disallowed during execution of a buy metaorder. This case requires a derivative-free optimization method. By using a direct-search method, namely the Generating Set Search (GSS) method, we find positive expected execution costs and *sparse* optimal strategies, *i.e.* it is optimal to trade with a few bursts at a high trading rate interspersed with long periods of no-trading.

In order to eliminate negative cost solutions, we propose two ways of regularizing the model, one based on the addition of a spread cost and one based on a modification of the instantaneous impact function. In the latter case the function becomes convex for sufficiently high trading rates. Both methods succeed in avoiding solution with negative costs and obviously reflect features of real markets.

This chapter is organized as follows. In Section 8.2 we present our results on the SQP and direct-search brute force minimization of the cost function. Section 8.3 presents the two proposed regularization methods in the context of the propagator model. In Appendix 8.4.1 we show the costs relative to a specific sub-optimal strategy, i.e. a sequence of bursts of trading. Our interest in such strategy is motivated by the numerical results of the brute force optimization.

8.2 Numerical optimization

In this section we study the cost optimization problem in the class of general piecewise constant strategies. In this case *general* means that these strategies are not necessarily deformations of continuous functions of time, such as those that can be obtained with the homotopy method. For this purpose we discretize the interval $[0, T]$ exactly as in Section 7.3.2 and minimize the cost function (7.48) with respect to the trading rates v_i subject to the constraint that the total quantity traded should be X. Recall that when the instantaneous impact function is linear, *i.e.* $f(v) \propto v$, the cost reduces to a N-dimensional quadratic form. In the nonlinear case, the numerical minimization involves finding the extrema of a complicated nonlinear function of N variables. We begin in Section 8.2.1 by presenting a motivating toy example that demonstrates how a nonlinear instantaneous market impact function f can lead to transaction-triggered price manipulation, *i.e.* optimal strategies where it is optimal to sell in some subintervals during a buy program. In the following subsection we consider the general case.

8.2.1 A motivating example

Here we show that with a nonlinear market impact function, the convexity condition on the decay kernel $G(\tau)$ is not sufficient to ensure the absence of transaction-triggered price manipulations for all the values $\gamma + \delta \geq 1$, *i.e.* the no-dynamic-arbitrage parametric zone. We illustrate this concept in the simplest case with $N = 2$

where the execution consists of two interval VWAPs. Setting $T = 1$, the expected liquidation cost (7.48) becomes

$$C\left[v_1, v_2\right] = \frac{1}{(1-\gamma)(2-\gamma)} \left(\frac{1}{2}\right)^{2-\gamma} \left[v_1 f\left(v_1\right) + v_2 f\left(v_2\right) + \left(2^{2-\gamma} - 2\right) v_2 f\left(v_1\right)\right]. \tag{8.1}$$

When $f(v) \propto v$, (8.1) reduces to the formula for a paraboloid in two dimensions. Imposing the constraint $v_1 + v_2 = 2X$, the problem reduces to a minimization with respect to v_1.

In Figure 8.1, with $\gamma = 0.5$, $X = 0.1$, we plot $C\left[v_1, v_2\right]$ against v_1 for various values of δ. We observe that for $\delta < 1$ the cost function has two local minima, one for a positive value of v_1 and one for a negative value of v_1. When $\delta \gtrsim 0.56$ the global minimum is the one with $v_1 > 0$, while for $\delta \lesssim 0.56$ the global minimum is attained for a negative value of v_1. It follows that if the impact function is strongly nonlinear, we can decrease the expected cost of a buy program with an intermediate sell trade, *i.e.* there is transaction-triggered price manipulation. Notice that, if we impose $v_1 \geq 0$, we have a boundary solution, *i.e.* it is optimal not to trade in the first interval, trading the whole order in the second interval.

We observe this behavior for all values of the parameters near the boundary of the no-dynamic-arbitrage region $\gamma + \delta \geq 1$, *i.e.* in the strong nonlinear region. This simple example also demonstrates that in the strong nonlinear regime, in the case of a buy program, it is better to sell during the first half of the trading period and then buy during the second half. In the next section we show numerically that this effect is accentuated when N is large.

8.2.2 Numerical cost minimization

In the general case we perform a non-convex constrained optimization of the cost function defined by the N variables v_i. The linear market impact case, *i.e.* $\delta = 1$, is a convex optimization problem, while in the non linear case the cost function is not convex and therefore we need to resort to numerical methods. We make use of two incomplete methods, *i.e.* we do not reach the global minimum with certainty. The first method is based on the Sequential Quadratic Programming (SQP) algorithm. This is one of the most successful methods for the numerical solution of constrained nonlinear optimization problems (NLP). It is an iterative procedure which models the NLP for a given iterate by a quadratic programming (QP) sub-problem, and then uses the solution to construct a new iterate. Convergence to a local minimum is then guaranteed. We use the routines implemented in Matlab [3, 2, 4].

When we search for strategies with the constraint that all trades should have the same sign, an algorithm based on derivatives, like SQP, can fail on the hyperplanes of the state space defined by $v_i = 0$, where the derivatives of the cost function diverge. For this case, we employ a second method based on a direct search approach, the *generating set search* (GSS) algorithm [150, 149], implemented in Matlab. This method does not require the computation of the derivatives of the cost function, because it searches directly directions of space where the cost decreases. When the

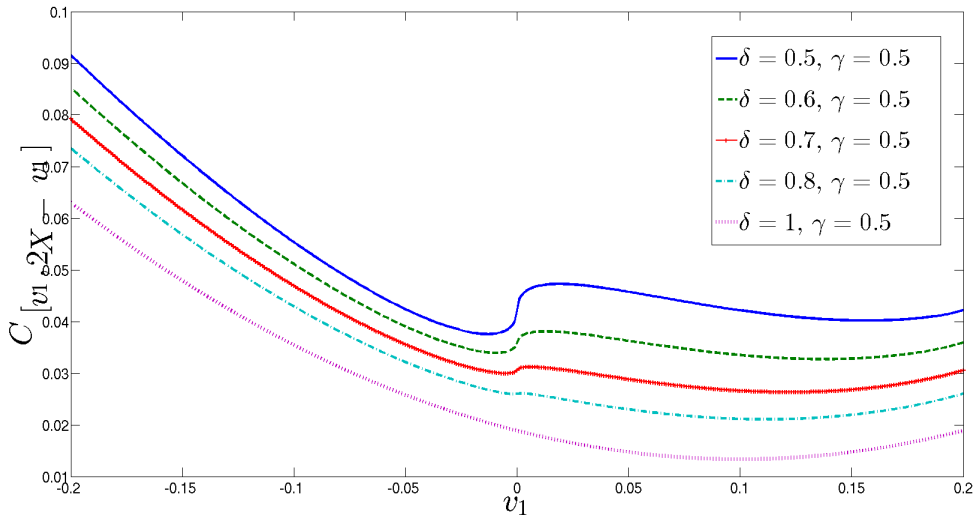

Figure 8.1: The cost function $C[v_1, 2X - v_1]$ for $X = 0.1, \gamma = 0.5$. The blue graph refers to $\delta = 0.5$, green to $\delta = 0.6$, red to $\delta = 0.7$, cyan to $\delta = 0.8$ and purple to $\delta = 1$ where the minimum is given by $v_1 = X$. In the nonlinear case there are two local minima.

search is close to the boundary of the feasible region, the set of search directions must include directions that conform to the geometry of the boundary. As would be expected, the GSS algorithm is more computationally expensive than the SQP algorithm.

We adopt a multiple random start approach, consisting of picking random starting points on the hyperplane defined by equation (7.42) (the constraint on the total quantity traded) and performing local SQP and direct search optimizations starting from these points. We then study the difference between the various extreme points, investigating whether they are local minima, and then among these extrema we select the solution with the smallest cost. Again, there is no guarantee that this corresponds to the global minimum.

8.2.3 Results

We tried different distributions for the initial points (*e.g.* a Dirichlet distribution on the hyperplane with various parameters), but we found qualitatively similar results in all cases. We illustrate here detailed results of such a procedure for $N = 100$ and $\gamma = \delta = 0.5$, $\gamma = 0.45, \delta = 0.55$ using 1000 starting points distributed uniformly on the hyper-plane $\sum_{i=1}^{N} v_i$ =const. This numerical analysis highlights three main features of the expected cost function (7.48) subject to the trading-volume constraint (7.42): (i) it has many local minima, (ii) there are many minima for which some $v_i < 0$, and (iii) the optimal cost can assume negative values for a buy-program. We

	DHAM	SQP	DIRECT	DHAM	SQP	DIRECT
δ	$\gamma = 0.45$	$\gamma = 0.45$	$\gamma = 0.45$	$\gamma = 0.5$	$\gamma = 0.5$	$\gamma = 0.5$
1.0	0.0116	**0.0115**	0.0115	0.0131	**0.0131**	0.0131
0.95	0.0130	**0.0128**	0.0129	0.0148	**0.0147**	0.0147
0.90	0.0143	**0.0136**	0.0140	0.0164	**0.0158**	0.0162
0.85	0.0162	**0.0139**	0.0151	0.0185	**0.0166**	0.0176
0.80	0.0179	**0.0138**	0.0162	0.0204	**0.0170**	0.0188
0.75	0.0198	**0.0132**	0.0169	0.0227	**0.0169**	0.0202
0.70	0.0218	**0.0117**	0.0184	0.0249	**0.0163**	0.0220
0.65	0.0235	**0.0092**	0.0191	0.0274	**0.0146**	0.0238
0.60	0.0251	**0.0047**	0.0201	0.0297	**0.0120**	0.0245
0.55	0.0275	**−0.0029**	0.0212	0.0323	**0.0075**	0.0262
0.50				0.0347	**0.0003**	0.0278

Table 8.1: Costs of three different strategies, DHAM, SQP and Direct-search in the no-dynamic-arbitrage region for $\gamma = 0.45, 0.5$. The numbers in boldface indicate strategies achieving the lowest expected cost. The difference between expected costs increases with the degree of non-linearity. In each case we use a GSS initial guess to obtain the DHAM solution. We use a discretization of $N = 100$ subintervals and the SQP and Direct-search optimization are performed by using 1000 starting points.

consider first the properties of the solution with the minimal cost among those found with our numerical method and then we investigate the properties of the landscape and the characterization of the suboptimal solutions. Lastly, in Section 8.2.3, we describe the results of optimization with the additional constraint that $v_i \geq 0$ (for a buy metaorder). We show that in this case, the optimal trading strategy is *sparse* when the model is strongly nonlinear, *i.e.* it is optimal to trade in few intense bursts.

Minimal cost solution

In Figure 8.2, we plot the optimal trading profile for a buy program corresponding to the global minimum of the cost function using our numerical minimization procedure. We observe that the optimal quantity to trade in each subinterval varies very irregularly with time. The cost minimizing solution consists of a series of bursts of intense but short-lived buying, separated by long periods when it is optimal to sell slowly. Evidently, the optimal solution admits transaction-triggered manipulation, as already shown in the toy $N = 2$ case analyzed above.

Strategies that are close-to-optimal are qualitatively similar, but the positions of their spikes can be very different. As an example, in Figure 8.3, we plot the four lowest cost solutions when $\gamma = 0.5$ and $\delta = 0.5$. All are characterized by a few intense positive spikes, separated by periods of slow selling. But these solutions differ in the positions of the bursts of buying. To quantify the differences between these solutions, we compute the Euclidian distance between them, including a simple VWAP strategy for reference. The resulting distance matrix is

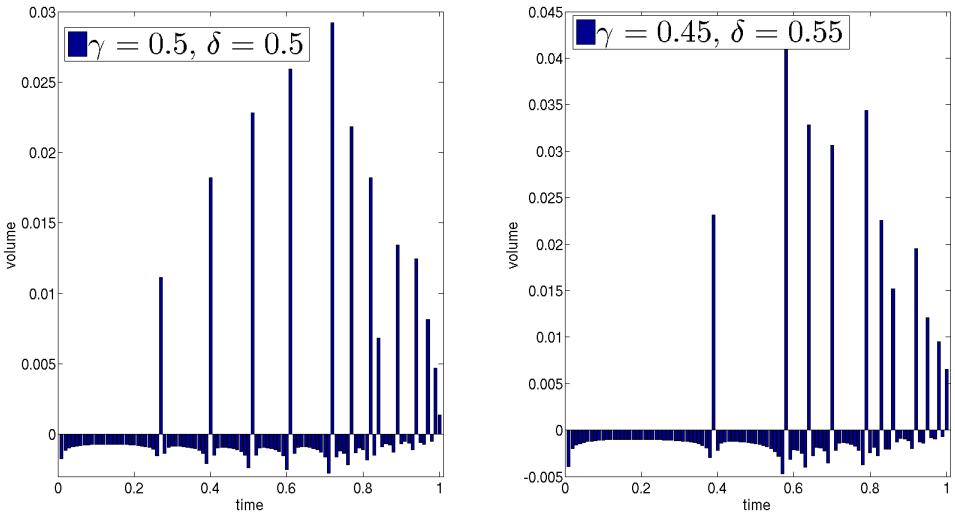

Figure 8.2: Optimal solution given by the SQP-algorithm for a buy-program where $X = 0.1$, *i.e.* 10% of a unitary market volume. We report the volume to be traded in each interval of time, *i.e.* $v_i\, T/N$.

$$D = \begin{pmatrix} 0 & 0.0780 & 0.0753 & 0.0890 & 0.0617 \\ 0.0780 & 0 & 0.0757 & 0.0757 & 0.0582 \\ 0.0753 & 0.0757 & 0 & 0.0799 & 0.0581 \\ 0.0890 & 0.0757 & 0.0799 & 0 & 0.0576 \\ 0.0617 & 0.0582 & 0.0581 & 0.0576 & 0 \end{pmatrix},$$

where the fifth row and column refer to the VWAP. We notice that all distances are quite similar. Strikingly however, the distance between any two SQP strategies in Figure 8.3 is greater than the distance of either of these to a simple VWAP; there is typically a multiplicity of strategies with bursts at different times, but with similar expected costs.

In Table 8.1 we report the costs for these various cost-minimizing solutions. We observe that the expected costs of SQP solutions are close to zero and even negative in the case $\gamma = 0.45$, $\delta = 0.55$. To further explore this undesirable behavior, in Figure 8.4 we plot the expected minimal cost for different values of γ as a function of δ and $N = 100$ (top panel) and $N = 150$ (bottom panel). First, we observe that costs are not a monotonic function of δ. Second, when $N = 100$ for $\gamma < 0.5$ we find a regime of δ for which the expected cost is negative. This effect becomes stronger when we increase the discretization. In fact, when $N = 150$ even for $\gamma = 0.5$ it is possible to find values of δ for which the minimal cost is negative.

A negative expected cost is a signature of the possibility of price manipulation. Indeed if market impact is entirely transient, with no permanent component, as in the model under consideration, it would then be possible to devise a round trip

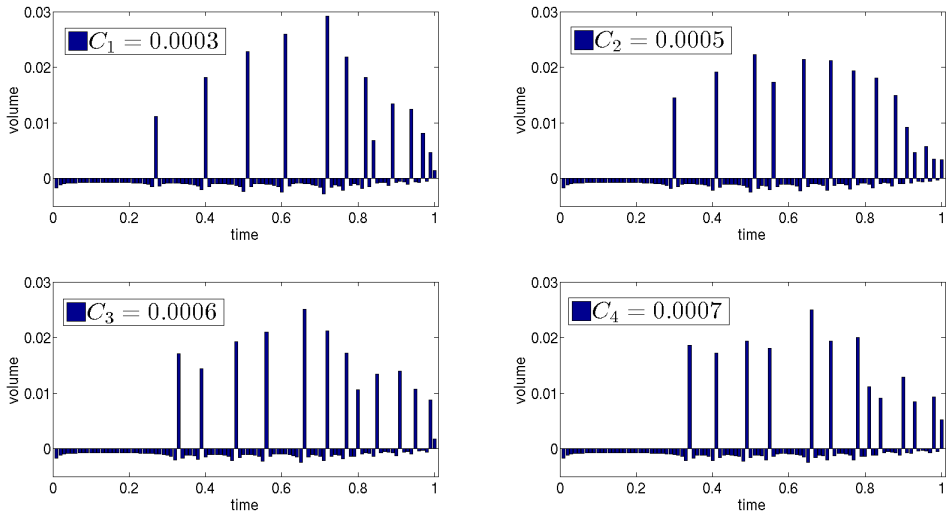

Figure 8.3: The four lowest cost solutions given by the SQP-algorithm for a buy-program where $X = 0.1$, *i.e.* 10% of a unitary market volume for $\gamma = 0.5$, $\delta = 0.5$. We report the volume to be traded in each interval of time, *i.e.* $v_i T/N$. The costs are reported in the insets.

strategy with positive (expected) revenue, *i.e.* a wonderful money machine. We cannot then escape the conclusion that a nonlinear transient impact model (7.1) with $f(v) = v^\delta$, $\delta < 1$ is mis-specified since it allows arbitrage opportunities.

Characterization of the cost landscape

We now investigate the structure of the cost landscape to be minimized and the properties of the extrema. We find a very large number of distinct extremal points. Jarque-Bera and chi-square goodness-of-fit tests cannot reject at the 5% significance level the Gaussian hypothesis for the distribution of cost of the local minima in the two data sets[4]. Thus we find a large number of extremal points which are quite similar in terms of expected cost. As expected the standard deviation of the cost of the local minima decreases as the market impact function becomes more linear. Similarly, the Euclidean distance between the local minima decreases when the model is closer to linear.

In general, the presence of many minima could reflect numerical error or a landscape that truly features multiple minima. The first case occurs frequently when the landscape is *sloppy* [52, 245], *i.e.* when there are directions of the space of variables along which the cost is substantially unchanged. A well known example of such a

[4]The Jarque-Bera test gives a p-value= 0.22 for the case $\gamma = \delta = 0.5$ and a p-value= 0.49 for the case $\gamma = 0.45$, $\delta = 0.55$. The chi-square goodness-of-fit test gives a p-value= 0.36 for the case $\gamma = \delta = 0.5$ and a p-value= 0.93 for the case $\gamma = 0.45$, $\delta = 0.55$.

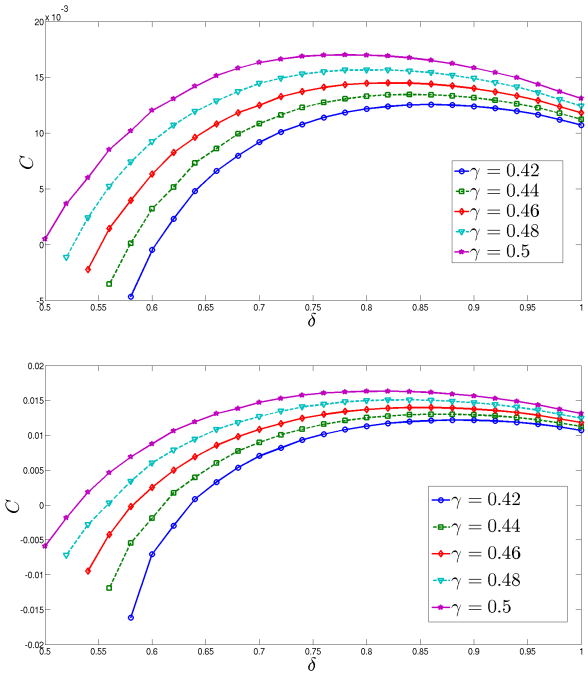

Figure 8.4: Optimal cost for the nonlinear transient impact model obtained with the SQP minimization. We consider a buy-program where $X = 0.1$, *i.e.* 10% of a unitary market volume and a discretization of $N = 100$ (top) and $N = 150$ (bottom) subintervals. We use 1000 starting points for each optimization. We consider only parameter in the no-dynamic-arbitrage zone. Holding γ fixed, the cost relative to the global minima is not a monotonic function of δ.

behavior is given by the Rosenbrock function, in which the global minimum is located inside a long narrow and parabolic shaped valley. Sloppiness can imply difficulties in finding the global minimum, because there is a manifold where the cost function is almost flat. Alternatively, the landscape may be *rugged*, with many local minima separated by local peaks. Rugged landscapes have attracted attention in physics [247], evolutionary biology [248], and computer science. An example of a rugged landscape is Kauffman's N-k fitness landscape [248]. The distribution of locally minimal energies of its spin-glass Hamiltonian is described by a normal distribution. In order to discriminate between these two alternatives and to characterize the cost landscape we performed two types of analysis.

First, we apply the second order condition to test which of the found extrema are actually minima and not, for example, saddle points. We perform a second derivative test on the local minima found by the SQP algorithm. In the case of constrained optimization the second-order sufficient condition for a minimum can be expressed in a determinant form of the bordered Hessian [60, 179]. We report the details of this analysis for the concave impact case when we choose a discretization $N = 100$. We have repeated this analysis for the concave-convex impact (see below) case finding

qualitatively the same results. Our optimization program starts from $1,000$ initial points, so we compute the above test on the final points where the SQP algorithm converges. For any value of the parameters δ and γ in the no-arbitrage region, more than 95% of the extremal points are actual minima. The remaining points are saddle points.

Second, we directly test the hypothesis that the landscape is sloppy, by using the eigenvalues and eigenvectors of the bordered Hessian of the lagrangian landscape function to identify its *stiff* and *sloppy* directions. One can study the sensitivity of the landscape function to changes by the eigenvalue spectra of the Hessian computed at a local minimum. Sloppy models are described by a constant logarithmic density for eigenvalues over six or more order of magnitude [52, 245]. The sensitivity of the landscape to changes is given by the square root of the eigenvalue. For sloppy models this means that one should move along the sloppiest eigen-direction a thousand times more than along the stiffest eigen-direction in order to change the function by the same amount. We have computed eigenvalue spectra of the bordered Hessian of the lagrangian function in the case of a concave market impact function and we found that the spectra are not compatible with a sloppy landscape. The bulk of the eigenvalues have similar small values of the order of 10^{-2}. This implies that the region near a local minimum is not flat.

In summary, the above analyses indicate that the vast majority of extremal points are actual minima and the region around a local minimum is not flat shaped in some direction. This suggests that the cost landscape under concave instantaneous market impact is rugged rather than sloppy. This situation is reminiscent of the search for the global minimum of the free-energy force field describing a protein [242]. The similarity with our problem is given by the structure of the free-energy, described by the sum of many Lennard-Jones potentials, *i.e.* a non-convex and non-periodic function. The great number of local minima is a consequence of summing non-convex functions and not a consequence of the presence of periodic functions. Similarly, in the concave impact case the cost is the sum of functions like the one described in Figure 8.1 for our 2-D toy model; such functions are non-convex and non-periodic.

Monotone strategies

In this section we consider monotone strategies, *i.e.* strategies where the position $X(t)$ is a non decreasing (non increasing) function of time for buy (sell) metaorders. We thus impose the absence of transaction-triggered manipulation. From the point of view of the numerical cost minimization, we impose the additional constraints $v_i \geq 0$, $\forall i$ (for a buy program). This additional constraint is analogous to the no-short-selling constraint in portfolio optimization [49].

We use a direct search method, specifically the generating set search algorithm [149]. Obviously, the cost is positive in this case; we report the costs of optimal strategies in Table 8.1. We find that the expected cost increases as the instantaneous market impact function becomes more nonlinear (*i.e.* as δ decreases). Moreover, expected costs are higher than that found by the SQP algorithm, but still significantly lower than expected of the strategies obtained using the DHAM approach. The

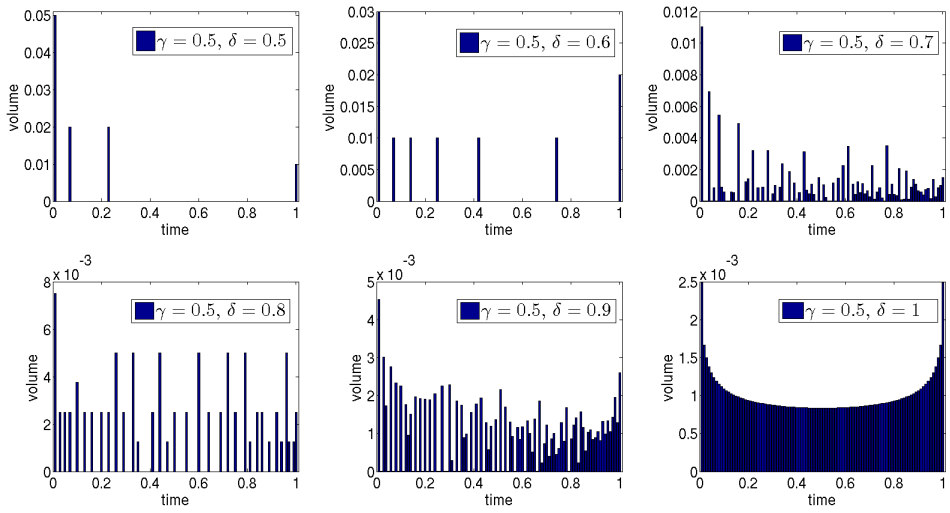

Figure 8.5: Optimal solutions given by the direct-search algorithm for a monotone buy-program where $X = 0.1$, *i.e.* 10% of a unitary market volume for $\gamma = 0.5$. We report the volume to be traded in each interval of time, *i.e.* $v_i T/N$.

structure of the optimal solutions is given in Figure 8.5 for the case $\gamma = 0.5$; results for $\gamma = 0.45$ are similar. The main feature that we observe is that as δ decreases, so that the market impact function becomes more nonlinear, the optimal trading profile becomes more and more *sparse*. The optimal strategy then consists of a few bursts of buying interspersed with long periods of no-trading. The geometrical interpretation of this result is that the solution lies on the boundary of a $(N-1)$-simplex. This is consistent with our unconstrained SQP results, where optimal strategies lie just beyond the edge of the simplex. The direct search simply stops at the edge of the feasible region described by the simplex. Finally notice that under the monotonicity constraint, it is optimal to start trading in the first interval, whilst without this constraint, it is optimal to slowly push down the price by selling before trading the first burst.

8.3 Regularizing the solution

In this section, we show that the nonlinear impact model (7.2) may be regularized using two different approaches, both of them reflecting important features of the market that we have so far neglected. In Section 8.3.1, we add a spread cost to the model in order to penalize wrong-way trading that may give rise to negative execution costs. This is equivalent to an L_1 or LASSO regularization. In Section 8.3.2 we modify the shape of the instantaneous market impact function f for high trading rates; the resulting market impact function is then concave-convex.

8.3.1 Adding a spread cost

In this section we add a spread cost to the model of equation (7.1) to penalize wrong-way trading. Equation (7.1) could be regarded as describing the evolution of the mid-price. When a market order is executed[5], there is an extra cost of half of the bid-ask spread $2\delta_S$, and the trading price is given by

$$S(t) = S(0) + \int_0^t f(\dot{x}(t)) G(t-s)\, ds + \int_0^t \sigma\, dW(s) + \delta_S \int_0^t \delta(s-t)\, sign(\dot{x}(s))\, ds, \tag{8.2}$$

The spread term is a temporary impact term that can be described by a δ-impact function; it affects only the price at which the order is executed and does not affect the market price. This term can represent also any cost or fee proportional to the absolute volume executed. The expected execution cost is then given by

$$C[\Pi] = C_G[\Pi] + C_S[\Pi] = \int_0^T \dot{x}(t) \int_0^t f(\dot{x}(s)) G(t-s)\, ds\, dt + \delta_S \int_0^T |\dot{x}(t)|dt. \tag{8.3}$$

The second term thus penalizes any strategy that consists of both buy and sell trades. This term is minimized, *i.e.* $C_S[\Pi] = \delta_S X$, for strategies where all trades have the same sign. This penalty is a form of L_1 or LASSO regularization widely used in computer science and used by Brodie et al. [49] to penalize short positions in Markowitz portfolio optimization. Busseti and Lillo [54] have calibrated optimal execution strategies on real data by regularizing them with a spread cost of this form.

In order to parameterize the relative cost of spread and impact, we define the dimensionless quantity $r = C_S[VWAP]/C_G[VWAP]$, which is the ratio between the spread cost and the impact cost for a VWAP execution according to the model of equation (8.2). This quantity is given by

$$\delta_S = r\, f(X) \frac{T^{1-\gamma}}{(1-\gamma)(2-\gamma)} \tag{8.4}$$

By choosing a value of r we set the value of δ_S. We consider the case of small spread cost, *i.e.* $r = 10\%$, and the case of high spread cost, *i.e.* $r = 50\%$ and perform a numerical optimization of the discretized cost function

$$C[v] = \sum_{i=1}^{N} \sum_{j=1}^{N} v_i\, f(v_j)\, A_{ij} + \delta_S \frac{T}{N} \sum_{i=1}^{N} |v_i|, \tag{8.5}$$

with the same parameters as in Section 8.2, $N = 100$, $X = 0.1$, for the case $\gamma = 0.45$, $\delta = 0.55$. Figure 8.6 plots the strategies corresponding to the global minimum in the numerical optimization in high spread cost and low spread cost cases respectively.

[5]Even if execution is with limit orders, one finds in practice that due to adverse selection, there is an extra cost of a portion of the bid-ask spread.

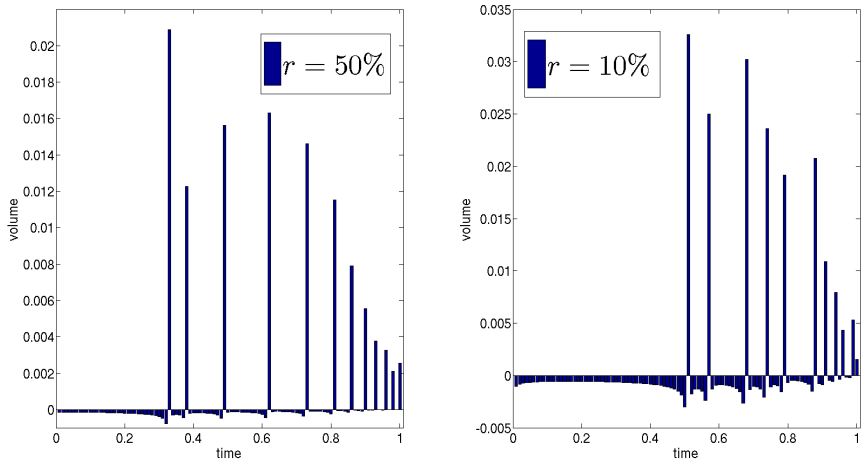

Figure 8.6: Optimal solution given by the SQP algorithm for a buy-program where $X = 0.1$, *i.e.* 10% of market volume, in presence of a spread cost. We report the volume to be traded in each interval of time, *i.e.* $v_i T/N$ for the case $\gamma = 0.45, \delta = 0.55$. On the left is the case of high spread cost, *i.e.* $r = 50\%$, on the right is the case of low spread cost, *i.e.* $r = 10\%$. The expected execution cost is $C_{SQP} = 0.026$ for $r = 50\%$ and $C_{SQP} = 5.9 \times 10^{-3}$ for $r = 10\%$.

Despite the fact that transaction-triggered price manipulation is still observed, in both cases the liquidation cost is now positive. In fact, we find $C_{SQP} = 0.026$ for $r = 50\%$ and $C_{SQP} = 5.9 \times 10^{-3}$ for $r = 10\%$. SQP strategies still outperform VWAP, GSS and DHAM strategies.

Finally, the effect of a high spread cost is a substantial reduction in negative trading volume similar to what is observed by Brodie et al. [49] for Markowitz portfolio weights. In that case a penalization term proportional to the sum of absolute values of weights leads to optimal solutions where the resulting portfolio is sparse with few assets and no-short positions. Similarly, in a nonlinear market impact model with a high spread-cost, optimal solutions for a buy-program appears to be characterized by a few bursts of trading-activity separated by intervals of time in which we have a very weak negative trading activity. Moreover, a high spread cost leads to liquidation costs similar to those corresponding to monotone strategies.

8.3.2 Concave-convex impact

Up to now, we have assumed that it is possible to trade arbitrarily fast whilst preserving the same functional form of the impact $f(v)$. Although this simplifies the mathematical treatment of the problem, it is unrealistic since we cannot trade at an arbitrarily high rate in practice. At some point the rate of trading is so high that one trades deeply into the order book, where less liquidity is available, and at this point one can expect that $f(\cdot)$ becomes convex. This motivates us to regularize the

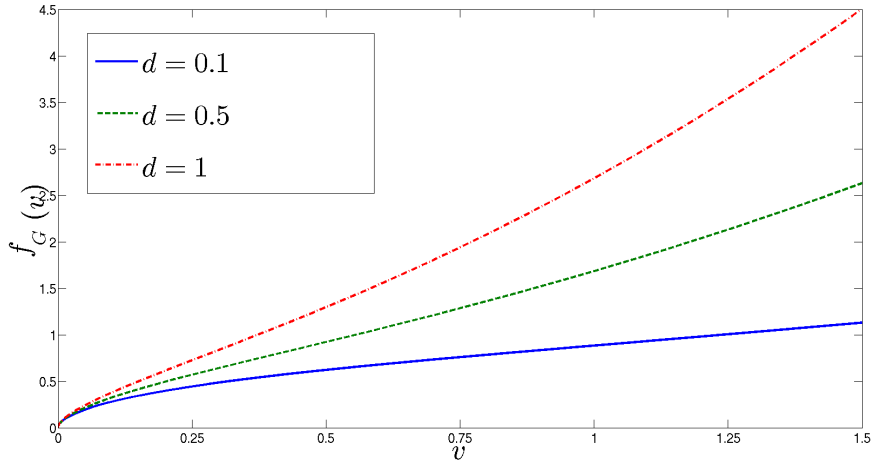

Figure 8.7: Concave-convex impact function for values of parameters: $c = 1$, $\delta = 0.55$, $X_M = 1$, $T = 1$.

transient impact model by postulating a concave-convex impact function that might both penalize excessively high trading rates and potentially eliminate the problem of negative expected liquidation costs. Specifically, we consider the following form for the impact function:

$$f_G(v) = c\ sign(v) \left\{ \left(\frac{|v|}{|v| + V} \right)^{\delta} + d \frac{|v|(|v| + V)}{V^2} \right\}, \tag{8.6}$$

where $V = X_M/T$ is the market volume X_M per unit time and c and d are positive constants. Figure 8.7 shows the form of this impact function. The constant d is a measure of the magnitude of the convex term with respect to the concave term, which is of order one. If $d \ll 1$ we recover a concave impact function. The parameter d sets directly also the value v^* where the function's convexity changes, where $f''_G(v^*) = 0$, see second column of Table 8.2. One could think of the rate v^* as approximating the maximum rate at which it is reasonable to trade.

We illustrate the results of numerical minimization of expected cost in the case $\gamma = 0.45, \delta = 0.55, N = 100$ for four different values of d: 0.1, 0.5, 1, 2, using 1000 starting points for each optimization. In Fig. 8.8, we plot the optimal strategies in each of these four cases. We observe that an increase in the magnitude of the convex term causes a decrease in maximum trading rates and an increase in the number of periods in which buying is optimal. The convex part of the impact acts like a barrier for high trading rates. In Table 8.2 we compare the costs of the SQP strategies with

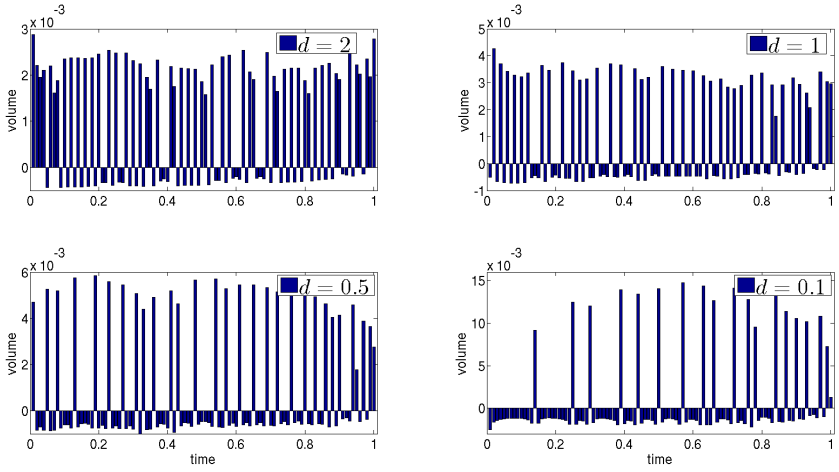

Figure 8.8: Optimal solution given by the SQP-algorithm for a buy-program where $X = 0.1$, *i.e.* 10% of market volume, in presence of a concave-convex impact. We report the volume to be traded in each interval of time, *i.e.* $v_i\, T/N$ for the case $\gamma = 0.45, \delta = 0.55, N = 100$. When the magnitude of the convex term increases, *i.e.* as d increases, trading rates decrease. Expected execution costs are given in Table 8.2.

respect with the cost of the corresponding VWAP strategy [6]. As the magnitude of the convex term increases we no longer find negative expected liquidation costs; at least in simulation, the presence of the convex part of the impact function is able to regularize the nonlinear transient impact model. It is worth noticing that the value of v^* is close to the mean trading rate in each case. Thus by knowing the impact function, one has a rough estimate of the optimal trading speed.

Finally, we extend our investigation to the whole no-dynamic-arbitrage region, in order to observe how the costs change. Figure 8.9 shows the expected cost for different values of γ as a function of δ, considering separately the case $d = 0.1$ and $d = 1$. In the former case, we observe a region of parameter space where the cost is negative, while if the convex term is sufficiently large, it is possible to eliminate negative costs in the strong nonlinear region. Further investigation shows that this result does not change as we increase N (data not shown). Finally we note that in the case of a concave-convex impact, as in the case of concave impact, the cost is a non-monotonic function of δ.

In conclusion, our findings emphasize the great importance that the specific shape of the instantaneous market impact function f has for the regularity of the optimal strategy.

[6] For the concave-convex impact function the cost of VWAP is $C_{VWAP} = \frac{c\,XT^{(1-\gamma)}}{(1-\gamma)(2-\gamma)} \left\{ \left(\frac{X}{X_M + X} \right)^{\delta} + d \left(\frac{X(X+X_M)}{(X_M)^2} \right) \right\}$.

d	v^*	$\langle v_{SQP} > 0 \rangle$	$\sigma\left(v_{SQP} > 0\right)$	Cost SQP	Cost VWAP
0.1	1.0755	1.1485	0.3193	-0.00245	0.03266
0.5	0.4256	0.4835	0.0911	0.01674	0.03782
1	0.2678	0.3229	0.0443	0.02887	0.04428
2	0.1639	0.2170	0.0292	0.04752	0.05718

Table 8.2: In the first column, we report the value of v^* for which $f''_G(v^*) = 0$. The other columns report data regarding the SQP optimization in the case $\gamma = 0.45$, $\delta = 0.55$, $N = 100$. The trading speed decreases when the magnitude of the convex term increases; the cost, instead, increases. The last column report the cost of a VWAP strategy in presence of a concave-convex impact.

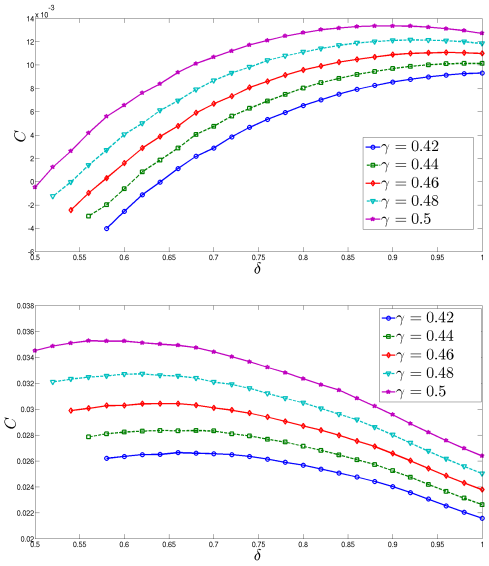

Figure 8.9: Cost of the optimal solution given by the SQP algorithm for a buy-program, where $X = 0.1$, *i.e.* 10% of a market volume and with $N = 100$ subintervals, and a concave-convex impact function with $d = 0.1$ (top) and $d = 1$ (bottom). We use 1000 starting points for each optimization and consider only parameters in the no-dynamic-arbitrage region defined by (7.6). We observe that holding γ fixed, expected cost is not a monotonic function of δ. For $d = 0.1$ negative costs are observed for any value of γ, while for $d = 1$ we observe no negative costs.

8.4 Appendices

8.4.1 Bursts of trading and waiting periods

The results of this chapter, regarding the optimal execution in presence of a non-linear market impact, lead to the conclusion that can be sub-optimal to trade by a burst strategy, i.e. a strategy where we have periods of no-trading. This intuition is strongly motivated by the results of section 8.2.3 regarding monotone strategies.

The constraint of positive trading rates implies irregular bursts strategies as optimal strategies. In this section we consider a subclass of such kind of strategies, imposing a regular shape for the trading rate.

In this appendix we consider the discretized environment developed in section 7.3.2. There we have considered the cost given by a piece-wise constant trading strategy. Here we consider a set of constant trading periods separated by periods of no-trading. We consider such simplified sub-optimal case in order to give some exact results when we consider strategies with waiting periods, i.e. periods of no-trading, in presence of a nonlinear market impact. The strategies studied in this chapter can be considered as a simplified version of the monotone strategies analyzed in section 8.2.3, where the transaction triggered price manipulations are not allowed.

Specifically, we consider a regular bursts strategy defined by a constant rate of trading on a regular subset of the overall period of execution $[0, T]$. It is much more simple than the complex strategies given by the SQP(or direct search) optimization. The main result is that such simplified regular strategy is able to decrease the expected cost respect to a VWAP execution.

The cost function

In this section we consider a generalization of the discretized cost of eq. 7.48 when we consider a general bursts strategy. Such strategy is defined by two parameters: the number of bursts N and the fraction $\alpha \in [0, 1]$ of the trading time T during which we trade. While for the time $(1 - \alpha) T$ we do not trade. Each burst lasts $\alpha T/N$ and is equispaced from the others. During each burst the trading speed is a constant $v_i, i \in \{1, 2, \cdots, N\}$. In the general case the speeds v_i can be different among each other. In this case we have the constraint given by the total traded volume X

$$\sum_{i=1}^{N} v_i = \frac{NX}{\alpha T}. \tag{8.7}$$

It is important to stress that here we are not solving the Urysohn equation 7.11, but only computing the expected cost for burst strategy. The expected cost is given by

$$C[v] = \frac{T^{2-\gamma}}{(1-\gamma)(2-\gamma)} \sum_{i=1}^{N} \sum_{j=1}^{N} v_i f(v_j) \tilde{A}_{ij}(\alpha), \tag{8.8}$$

where $\tilde{A}(\alpha)$ is a Toeplitz matrix that describes the decay kernel $G(t-s) = (t-s)^{-\gamma}$

$$\begin{aligned}
\tilde{A}_{ij} &= 0;\ j > i; \\
\tilde{A}_{ii} &= \left(\frac{\alpha}{N}\right)^{2-\gamma}; \\
\tilde{A}_{ij} &= \left[\left(1 + (i-j+1-N)\frac{N-\alpha}{N(N-1)}\right)^{2-\gamma} - 2\left(\frac{N-\alpha}{N} + (i-j+1-N)\frac{N-\alpha}{N(N-1)}\right)^{2-\gamma}\right. \\
&\quad \left. + \left(\frac{N-2\alpha}{N} + (i-j+1-N)\frac{N-\alpha}{N(N-1)}\right)^{2-\gamma}\right];\ j \leq i.
\end{aligned} \tag{8.9}$$

In the case $\alpha = 1$ we recover the expressions given by eqs. 7.49, i.e. the case where trading is allowed on all the period $[0, T]$. In this chapter we consider only a regular bursts strategy, i.e. $v_i = v_c = X/(\alpha T), \forall i$. This choice gives a simplified version of the cost function, where it is possible to exploit the properties of the Toeplitz matrix $\tilde{A}_{ij} = \tilde{A}_{i-j}$. The cost for a regular bursts strategy is given by

$$C[v] = \frac{T^{2-\gamma}}{(1-\gamma)(2-\gamma)} v_c f(v_c) \sum_{i=1}^{N} i\tilde{I}_i, \tag{8.10}$$

where we have used the index transformation $N - (i-j) \to i$ to map $\tilde{A}_{ij}$ into $i\tilde{I}_i$

$$\begin{aligned}
\tilde{I}_N &= \left(\frac{\alpha}{N}\right)^{2-\gamma}; \\
\tilde{I}_i &= \left[\left(1 + (1-i)\frac{N-\alpha}{N(N-1)}\right)^{2-\gamma} - 2\left(\frac{N-\alpha}{N} + (1-i)\frac{N-\alpha}{N(N-1)}\right)^{2-\gamma}\right. \\
&\quad \left. + \left(\frac{N-2\alpha}{N} + (1-i)\frac{N-\alpha}{N(N-1)}\right)^{2-\gamma}\right];\ i \leq N.
\end{aligned} \tag{8.11}$$

An example of the regular strategy is given in fig.8.10 for the case $X = 0.1, N = 10, \alpha = 0.5, T = 1$. Each burst has a duration $\alpha T/N$ and is separated by an interval ΔT in which no trading occurs. The waiting period between two bursts is defined by $\Delta = (1-\alpha)/(N-1)$. The total time of trading is αT. The interval $[0, T]$ is partitioned into $M = 2N - 1$ sub-intervals and the i-th left bound of each burst is given by:

$$t_i = (i-1)\frac{(N-\alpha)}{N(N-1)}T; \quad i = 1, \cdots, N; \quad N \geq 2. \tag{8.12}$$

Concave market impact

We make the following assumptions: $f(v) = c\left(\frac{v}{V}\right)^{\delta}$, $G = (t-s)^{-\gamma}$, where we consider the no-dynamic-arbitrage zone defined in chapter 7 by the eq. 7.6. We consider only positive trading speeds v. The differences with the concave impact function considered in chapter 7 regard only the constants c and V, that are introduced here in order to

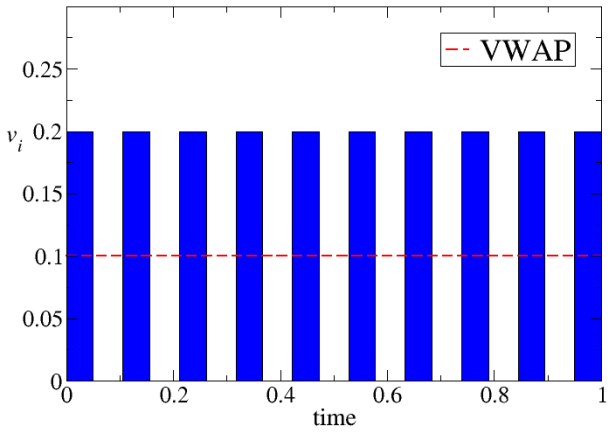

Figure 8.10: Regular bursts strategy for the case $X = 0.1, N = 10, \alpha = 0.5, T = 1$. The corresponding VWAP execution is reported.

make the impact function more realistic from a practitioner point of view [10]. The constant $V = X_M/T$ in f is the market volume X_M per unit time, i.e. a constant average trading rate of the all market. The benchmark model is the one with linear and permanent impact, i.e. $\delta = 1$ and $\gamma = 0$, whose optimal solution is the VWAP [15]. The square-root process corresponds to $\delta = \gamma = 1/2$. The expected cost of a VWAP, defined by a constant trading speed $v = X/T$, is

$$C^{[VWAP]} = c\frac{1}{(1-\gamma)(2-\gamma)} X \left(\frac{X}{V}\right)^{\delta} T^{1-\gamma-\delta}, \tag{8.13}$$

and the impact of the metaorder, given by eq. 7.4 for $t = T$, is

$$I = S(T) - S(0) = \frac{c}{1-\gamma}\left(\frac{X}{V}\right)^{\delta} T^{1-\gamma-\delta}, \tag{8.14}$$

thus the exponent describing the dependence of the impact from the volume of the metaorder is the same as the one describing the impact of individual trades. Moreover, it is worth noticing that both the impact and the cost are independent of the time of execution $T \leq T_d$ if $\gamma + \delta = 1$. This critical parametric zone is actually the lower border of the no-dynamic-arbitrage zone. On the other hand if $\gamma + \delta > 1$ it is more convenient to trade in a longer time period, because for $T \to \infty$ the cost of the VWAP execution goes to zero.

The expected cost C of a regular bursts strategy, defined by bursts of equal speed $v_c = X/(\alpha T)$, can be connected to the cost of a VWAP strategy of duration T

$$C[v] = c\,X\left(\frac{X}{V}\right)^{\delta} \frac{1}{(1-\gamma)(2-\gamma)} \frac{T^{1-\gamma-\delta}}{\alpha^{\delta+1}} \sum_{i=1}^{N} i\tilde{I}_i = C^{[VWAP]}\alpha^{-\delta-1} \sum_{i=1}^{N} i\tilde{I}_i. \tag{8.15}$$

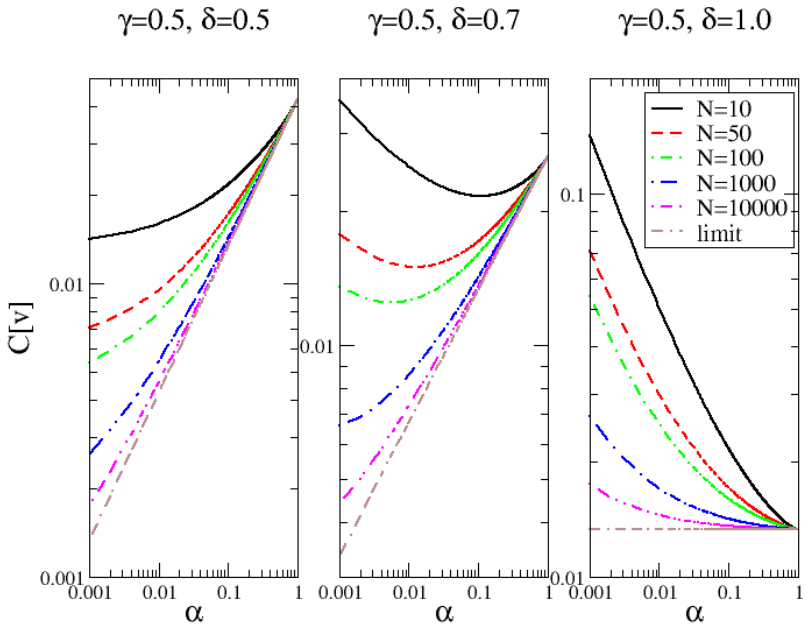

Figure 8.11: Cost of the regular bursts strategy for the case $X = 0.1, T = 1$. All costs converge to the cost of the VWAP execution for $\alpha = 1$. The limit cost is given by eq. 8.17.

For $\alpha = 1$, the sum is equal to 1 and the cost coincides with the one of the VWAP. Let us keep the value of N fixed and let us take the limit of the cost for $\alpha \to 0$, which corresponds to very intense and short lived bursts. We have

$$\lim_{\alpha \to 0} C[v] = C^{[VWAP]} \times \begin{cases} +\infty & \gamma + \delta > 1 \\ N^{-\delta} & \gamma + \delta = 1 \quad \delta < 1 \\ 1 & \delta = 1, \quad \gamma = 0 \end{cases} \tag{8.16}$$

The fact that for $\gamma + \delta > 1$ the limit is infinite implies that, given the number of bursts N, there exists an optimal value of α (in the limit 1) for which the cost is minimized. Numerical investigations indicate that the position of the minimum for α becomes smaller and smaller when one increases the number N of bursts. On the other hand at critical state $\gamma + \delta = 1$ and given the number of bursts N, numerical investigations indicate that the minimum is achieved at $\alpha = 0$. Figure 8.11 shows the cost as a function of α for different values of N in the two regimes of $\delta + \gamma = 1$ and $\delta + \gamma > 1$. In the linear case, i.e. $\delta = 1$, the costs of bursts strategy are higher than that of VWAP execution.

We can understand the different behavior in the critical and non critical state in the following way. At the critical state the execution cost of a VWAP is independent on the execution time. Therefore it is more convenient to separate burst (which are locally VWAP of duration $\alpha T/N$) as much as possible to exploit the price decay

during the non trading periods. In the non critical case ($\delta+\gamma > 1$) it is still convenient to separate bursts, but by decreasing the duration of the bursts one increases the cost per burst, therefore it is better to have longer bursts, i.e. less separated. The tradeoff between these two opposite tensions explains the presence of an optimal non vanishing value of α.

Let us now keep α fixed and let us take the limit of the cost for $N \to \infty$, which corresponds to infinitely many bursts but very short in time. We obtain

$$\lim_{N\to\infty} C = c\,X \left(\frac{X}{V}\right)^{\delta} \frac{1}{(1-\gamma)\,(2-\gamma)} T^{1-\gamma-\delta}\alpha^{1-\delta} = C^{[VWAP]}\alpha^{1-\delta}. \tag{8.17}$$

This result is obtained by the result: $\lim_{N\to\infty} \sum_{i=1}^{N} i\tilde{I}_i = \alpha^2$. If we want to reduce the cost to 0 we have to decrease the trading time $\alpha\,T$ to 0. Notice that the vanishing of the cost is possible only if $\delta < 1$. For linear impact (with or without decay) it is impossible to drive the impact to zero. As expected, if we send $\alpha \to 1$ in eq. 8.17 we obtain again the VWAP cost. The two limits $\alpha \to 0$ and $N \to \infty$ do not commute in general. The two limits commute only on the critical region $\gamma + \delta = 1$ when $\delta < 1$.

However, in order to send the cost to zero we should send $\alpha \to 0$ and this is impossible for any real execution strategy. The point that we stress as conclusion is that the non-linearity of the market impact function, i.e. in this case a concave impact function $f(v) \propto v^{\delta}$, $\delta < 1$, allows for a reduction of the expected execution costs by trading in bursts. This results are consistent with the results found in sec. 8.2.3 by means of the direct search algorithm. The minimum of the cost function in the constrained case, i.e. $v_i \geq 0$, is located on the boundary of the simplex defined by eq. 8.7.

8.4.2 Second order condition for constrained optimization

The introduction of Lagrange multipliers as additional variables makes it possible to apply to the constrained-extremum problem the same first order condition used in the free-extremum problem. It is tempting to go a step further and borrow the second-order sufficient conditions as well. This, however should not be done. As in the case of free extremum, it is possible to express the second-order sufficient condition in a determinant form. In place of the Hessian, in the constrained case we shall encounter what is known as a *bordered Hessian* [60, 179]. This matrix is simply the Hessian of the Lagrangian function, where the borders are given by the first derivatives of the constraint function. We illustrate it in our specific case, where we have one linear constraint and a cost function defined on $\mathbb{R}^N$ by the equation (7.48). The second

derivatives of the Lagrangian are given by

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \lambda^2} &= 0, \\
\frac{\partial^2 \mathcal{L}}{\partial \lambda \partial v_k} &= -1, \\
\frac{\partial^2 \mathcal{L}}{\partial v_i \partial \lambda} &= -1, \\
\frac{\partial^2 \mathcal{L}}{\partial v_i \partial v_k} &= f'(v_k) A_{ik} + f'(v_i) A_{ki} + \delta_{ik} \sum_{j=1}^{N} v_j f''(v_k) A_{jk},
\end{aligned} \tag{8.18}$$

where $i, k = 1, \cdots, N$ and the borders are given by the second derivative respect to λ and v_i, *i.e.* the derivatives of the constraint with respect to v_i.

The sufficient conditions imply the computation of the determinant of $N-1$ minors of the bordered Hessian. Let H_j be the upper left $j \times j$ submatrix of the bordered Hessian, we have the following sufficient conditions for a local stationary point

1. If $-\det(H_j) > 0$ for $3 < j < N+1$, the cost function has a local minimum.
2. If $(-1)^{j-1}\det(H_j) > 0$ for $3 < j < N+1$, the cost function has a local maximum.
3. If $\det(H_j) \neq 0$ for $3 < j < N+1$, but fall into a different pattern of signs than the above two cases, then the critical point is some point of saddle.
4. If some of the $\det(H_j) = 0$, the character of the critical point is indeterminate at the level of second order partial derivatives.

Part III

Appendix and Bibliography

Appendix A

Onset of chaotic dynamics in neural networks

A neural-network model is proposed as a testbed for the characterization of the chaotic dynamics emerging in a context where the coupling is, on the averge, neither excitatory nor inhibitory. The proposed discrete-time model generalizes within a single framework, two different setups previously studied in the literature. With the help of theoretical mean-field arguments and numerical simulations on GPUs, we characterize the transition and show that the chaotic dynamics is extensive (i.e. that the number of active degrees of freedom is proportional to the network size) from the very beginning. Besides the coupling strength, two parameters play a crucial role: (i) one controls the local dissipation and determines the shape of the initial part of the Lyapunov spectrum as well as the shape of the correlation function; (ii) the other, which corresponds to the amplitude of an effective random field, determines the nature of the transition.

A.1 Introduction

One of the relevant questions that arise in the context of neural networks is the emergence and self-sustainment of an irregular dynamics. The problem can be appreciated by referring to the *local mean field* (LMF), i.e. the effective field seen by each neuron, as a result of its interactions with the other neurons. The LMF is typically the sum of K terms, where K is the network connectivity (for the sake of simplicity we assume K to be constant across the network). Roughly speaking, three different mechanisms can be identified: (i) synchronization among a large subset of neurons; (ii) "statistical" fluctuations of the LMF; (ii) onset of a collective motion. The first one has been much studied in the literature [187, 22]: it essentially boils down to recognize that many degrees of freedom (the synchronized ones) are somehow eliminated, with the net result that the microscopic dynamics is low-dimensional. The second mechanism is a somehow trivial property of high-dimensional systems: it is a consequence of the "naturally" chaotic dynamics exhibited by nonlinear dynamical systems. As a result, the LMF is subject to "statistical" fluctuations of size $1/\sqrt{K}$. In this last

case, however, for large K, the amplitude of the LMF is dominated by its average and one expects that the amount of chaos induced by the coupling decreases upon increasing K. This is indeed the scenario observed in typical massively coupled phase oscillators, when they can behave chaotically only because of the coupling [196].

The third mechanism is based on a subtle form of synchronization: any two neurons subject to the same input field do not, strictly speaking, synchronize, but the averge response of an entire population of neurons reacts to the external modulation in a way that its fluctuations are self-sustained and a collective motion sets in. A striking example is discussed in [174], there the LMF exhibits an irregular evolution even though the "microscopic" dynamics is linearly stable. The difference between this and the previous mechanism is clearly proven by the fact that it survives even for arbitrarily large connectivities, when statistical fluctuations become negligible.

With reference to the second mechanism, the only case when the average does not dominate and fluctuations are negligible is when the average itself almost vanishes. This is indeed the setup first studied in [232], where it was introduced a simple, continuous-time, model describing the evolution of the coarse-grained neural activity and later implemented in the context of a globally coupled map [57]. In both cases, a critical value of the coupling strength has been found, beyond which a chaotic dynamics sets in. Later, the same idea has been extended to more realistic models [243] which describe the instantaneous spiking activity (see, e.g., leaky integrate-and-fire neurons): there, the nearly-vanishing average is the result of a balance between excitation and inhibition. Although *balanced* states have been repeatedly studied in the literature (see [144, 189] for a clean discussion thereof), it is still unclear under which conditions they can spontaneously arise. For this reason, it is still worth to study more abstract models, as they can help to better identify the underlying general mechanisms [86, 178].

In this paper we adopt this line of thought, proposing a single framework which encompasses the models considered in [232, 57] with the double goal of finding unifying elements (and differences) and to characterize the corresponding chaotic dynamics that was not explored in neither of the two seminal papers. Our model contains two important parameters (besides the coupling strength): an effective random field (absent in [232]) and a local contraction (maximal in [57]). With the help of a mean field approach borrowed from [57], we derive semi-analytical expressions for the "Lyapunov" spectra in the stable regime (where the dynamics converges towards a fixed point). As a result, we find that random matrix theory provides very good approximations (in some cases, exact expressions) for the maximum exponent, but also that the spectral shape differs substantially from the prediction based on the circle law [114].

Moreover, we carry out an extensive numerical study of the Lyapunov spectra above the transition, finding that the dynamics is extensive, i.e. the number of active degrees of freedom is proportional to the network size. This is analogous to standard spatially extended systems, although here this property is non trivial due to the non-additivity of the dynamics [173].

In section II we introduce the model and show the results of numerical simulations in some key cases. Sec. III is devoted to a mean field analysis, which allows deriving

approximate expressions for the fluctuations of the local field and for the Lyapunov exponents. In Sec. IV we determine the Lyapunov spectrum below the transition and show that, at variance with the claims made in [57], the mean-field prediction is not exact in the presence of an effective random field. In Sec. V we focus on the chaotic phase, that is characterized both with the help of Lyapunov spectra and of the correlation function of the local field. Finally, section VI is devoted to a discussion of the major results and of the open problems.

A.2 The model

Let us consider an ensemble of N neurons, each characterized by the continuous variable $-\infty < h_i < +\infty$. The model is defined by dimensionless variables, the evolution rule is

$$h_i(t+1) = \gamma h_i(t) + \frac{g}{\sqrt{N}} \sum_j J_{ij} \left[a + \tanh h_j(t) \right] , \tag{A.1}$$

where: (i) g is the coupling gauge; (ii) J_{ij}s are i.i.d. variables (referred to as the coupling constants) with zero average, unit variance, and a Gaussian distribution; (iii) a plays the role of a random external field, which breaks the homogeneity of the $h_i = 0$ stationary solution; (iv) $\gamma < 1$ accounts for the stability of the local field dynamics. The variable $u_j = \tanh h_j$ can be interpreted as the "activity" of the jth neuron: $u_j = -1$ and $u_j = 1$ correspond to the inactive and active state, respectively.

The sum in the r.h.s. is the LMF mentioned in the Introduction. In this setup, each neuron is coupled with all other neurons so that the connectivity coincides with the network size, i.e. $K = N$. By symmetry reasons, the average value of the LMF is zero and, if we assume that each term is of order $\mathcal{O}(1)$, the sum is of order $\mathcal{O}(\sqrt{N})$ and the LMF of order $\mathcal{O}(1)$. In other words, the statistical fluctuations do not vanish in the thermodynamic limit.

For $a = 0$ the model (A.1) reduces to the discrete-time version of the system studied in [232]: the continuous-time limit is attained for $\gamma \to 1$ and a finite value of the effective coupling constant

$$G = \frac{g}{1-\gamma} . \tag{A.2}$$

By setting $\gamma = 0$ and $a = 1$, we instead recover the model studied in [57]: notice, however, that the parameter g defined in this paper is half of that one introduced in [57]. Moreover, for the sake of completeness, we recall that a more general setup was also considered in [57], where the average of the coupling constant is slightly different from zero: as no substantial differences where observed, we leave this option out of our model.

Let us start by illustrating the dynamics for different parameter values. It is convenient to introduce two order parameters: (i) the ensemble variance

$$q_e^2(t) \equiv \overline{h_i^2}(t) \tag{A.3}$$

of the instantaneous fields h_i (by symmetry reasons, the mean value is equal to zero), where the overline denotes an ensemble average over all nodes; (ii) the temporal variance

$$q_t^2 \equiv \overline{\langle h_i^2 \rangle_t - \langle h_i \rangle_t^2} \tag{A.4}$$

of the local-field fluctuations, where $\langle \cdot \rangle_t$ denotes a time average.

As a general remark, in all regimes $q_e^2(t)$ is always constant, meaning that the model does not exhibit any collective dynamics in the thermodynamic limit (at variance with [174]). One can, nevertheless, distinguish three classes of behaviour: (FP0) a trivial fixed point $q_e = 0$, where all fields vanish; (FP1) where the fields h_i are different from each other and constant in time, so that $q_e > 0$; (CP) a microscopically chaotic phase, where $q_e > 0$ and $q_t > 0$.

We have focussed our studies on four different sets of parameters: (I) $\gamma = 1/2$, $a = 1/2$, as a typical instance of a generic point in parameter space; (II) $\gamma = 0$, $a = 1$, which, besides corresponding to a maximal contraction, coincides with the case studied in [57]; (III) $\gamma = 0.5$, $a = 0$, as a generic example with no external random field; (IV) $\gamma = 0.9$, $a = 0$, which is a discrete-time version of the model studied in [232].

In Fig. A.1 we have plotted q_e and q_t as a function of the coupling strength. In the two upper panels we see that q_e does not reveal any indication of a phase-transition: q_e increases smoothly over the entire range of g values. Upon monitoring q_t, we instead see that there exists a critical value g_c, above which the temporal variance is strictly larger than zero (see the lower sets of symbols in the two upper panels), revealing the existence of a microscopic dynamics.

In the absence of an external field ($a = 0$), the scenario is different. The transition can be detected directly by monitoring q_e, which becomes larger than zero as soon as $q_t > 0$. In fact, more than that, q_t is practically equal to q_e and the mutual differences vanish upon increasing the network size. In practice, we can speak of "ergodic" behaviour, since temporal and ensemble fluctuations coincide with one another, or, more properly, claim that all neurons exhibit the same behavior.

Altogether, in the presence of an external field, the diversity among the various neurons masks the onset of a dynamical behaviour (q_e is larger than zero both above and below the transition).

A.3 Mean field analysis

The setup (II) has been explored in [57] with the help of a mean-field approach, while a more sophisticated dynamical mean field has been developed to analyse the setup (IV) [232]. As the latter approach requires a continuous-time dynamics and the absence of an external field, here we adopt the former approach, which is approximate but rather accurate and can be easily implemented in all cases.

The starting point is the instantaneous probability distribution of the field h_i, for which we derive a self-consistent equation. Let us start considering the regimes FP0 and FP1. In this case one can drop the dependence on t and equate $h_i(t+1)(1-\gamma)$ to the coupling term. By further assuming that all addenda (and factors) in the sum

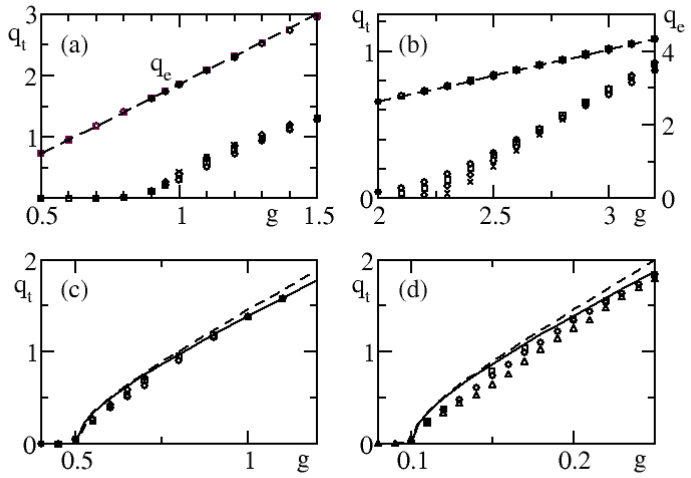

Figure A.1: Ensemble and temporal fluctuations in four points of the parameter space. (a-d) correspond to (I-IV). In all panels, triangles, circles, squares, diamonds, and crosses, correspond to, $N = 256$, 512, 1024, 2048, and 4096, respectively. In panel (b) the q_e scale is reported on the right hand side. The dashed lines correspond to the theoretical prediction from Eq. (A.5), while the solid lines correspond to Eq. (A.7). In the lower panels only the symbols referring to q_t are reported.

are mutually independent, one obtains

$$q_e^2 = G^2 \left\langle (a + \tanh q_e h)^2 \right\rangle = G^2 \Gamma^2(a, q_e), \tag{A.5}$$

where G is defined in (A.2), while the angular brackets without subscripts denote the Gaussian average

$$\langle F(h) \rangle = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dh F(h) \mathrm{e}^{-h^2/2} \,. \tag{A.6}$$

In the presence of a time-dependent dynamics, we expect the above equation to hold exactly for $\gamma = 0$, since the attraction term disappears in the r.h.s. (this hypothesis has been checked in [57] for $a = 1$). For $\gamma > 0$, $h_i(t+1)$ and $h_i(t)$ are only partially correlated and, we, thereby expect the above equation to provide just an approximate solution.

Let us now explore the implications of Eq. (A.5), starting from the case $a = 0$. One can easily check that since $\Gamma(0, q_e) = 0$, $q_e = 0$ is a valid solution for any value of G (see the lower curve in Fig. A.2). Since $[d\Gamma/dq](0, 0) = 1$, a second fixed point exists when $G > G_c = 1$. This is the transition observed in [232] in the limit of $\gamma \to 1$. The dashed lines in the lower panels of Fig. A.1 are the solutions of the self-consistent equation (A.5). There we see a good, though non perfect agreement with the direct numerical simulations.

For $a > 0$, there always exists one and only one solution $q_e = G\Gamma(a, q_e)$ no matter how small is a (see the upper curve in Fig. A.2). The dashed lines in the upper panels of Fig. A.1 correspond again to the theoretical predictions which are, in this case, essentially indistinguishable from the numerics.

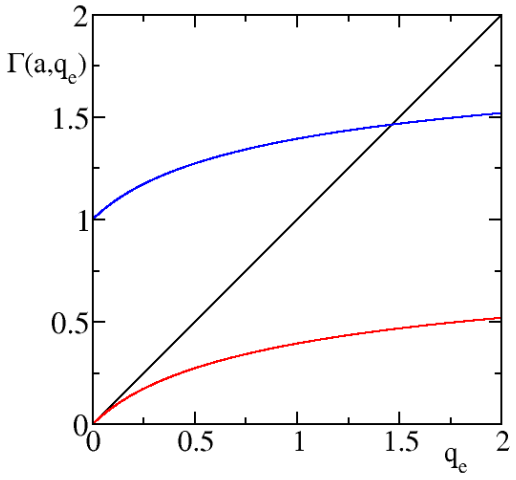

Figure A.2: (color online) Behaviour of the function $\Gamma(a, q_e)$ for $a = 0$ (lower curve) and $a = 1$ (upper curve).

No simple mean-field argument exists which can help to determine q_t. The toughness of the task is due to the fact that the macroscopic distribution does not exhibit any qualitative change in the presence of an external field.

In the limit case of a continuous dynamics (essentially case (IV)), one can develop a dynamical mean-field approach, where the argument of a self-consistent equation is the correlation function of the field. This approach eventually predicts that [232]

$$q_t^2 = 2G^2 \left[\left\langle (\log\cosh q_t h)^2 \right\rangle - \left(\langle \log\cosh q_t h \rangle \right)^2 \right] . \tag{A.7}$$

This formula is expressed with reference to q_t, but one can equivalently think of q_e, as the quantities concide for $a = 0$. The solid line in Fig. A.1d corresponds to this prediction, which indeed reproduces more accurately the numerical findings. Interestingly, the formula works perfectly well also for (III) (see panel c), although it is not a priori expected.

A.4 Linear stability

The most appropriate way to investigate the onset of a dynamical transition is by studying the evolution of infinitesimal perturbations $\delta\vec{h}(t)$. The corresponding equation is

$$\delta\vec{h}(t+1) = \mathbf{M}(\vec{h})\delta\vec{h}(t), \tag{A.8}$$

where

$$M_{ij} = \gamma \mathbf{I} + g L_{ij} \quad , \quad L_{ij} = \frac{1}{\sqrt{N}} \frac{J_{ij}}{\cosh^2 h_j} \,. \tag{A.9}$$

In the case of a fixed point, the stability is determined by the eigenvalues μ of the matrix $\mathbf{M}$ and, in particular, by the condition that the moduli of all eigenvalues are strictly smaller than 1. Here and in the following, we prefer, however, to make reference to the logarithms of the eigenvalues, $\lambda = \log|\mu|$, as they are analogous to the Lyapunov exponents of the chaotic dynamics and we indeed refer to them as to the Lyapunov exponents with a slight abuse of the notation.

In Fig. A.3 we plot the Lyapunov spectra numerically obtained below the transition in cases (I), (II), and (III). We indeed see that all the spectra lie below 0, confirming the stability of the fixed point solution. It is possible to go beyond numerical observations: for $a = 0$, the state $h_i = 0$ is a fixed point for any G value; since the J_{ij} are, by construction, i.i.d. elements with unit variance. Girko's theorem [114] implies that the eigenvalues of $\mathbf{L}$ are distributed inside a circle of radius $\sigma = 1$ centered around (0,0), while those of $\mathbf{M}$ are multiplied by g and shifted by γ, so that stability is lost when $\log(\gamma + g_c) > 0$. Therefore, the critical point is $g_c = 1 - \gamma$, consistently with the direct simulations reported in panels (c) and (d) of Fig. A.1.

More than that, the knowledge of eigenvalue distribution allows predicting the entire shape of the Lyapunov spectrum (in the stable regime). We start deriving an explicit expression for $\gamma = 0$ and $a = 0$. By denoting with m the modulus of the eigenvalue μ ($m = |\mu|$), the circular law implies that the distribution $P(m)$ of the moduli is $P(m) = 2m/g^2\sigma^2$ (for $m < \sigma g$). The Lyapunov spectrum is determined by expressing the value λ of the Lyapunov exponent versus the so-called integrated density ρ (defined in such a way that $\rho = 0$ corresponds to the maximum exponent). It is therefore convenient to pass from $P(m)$, to its integral

$$\rho(m) = \int_m^{g\sigma} dm P(m) = 1 - \frac{m^2}{(g\sigma)^2} \,. \tag{A.10}$$

By then recalling that $\lambda = \log m$,

$$\rho(\lambda) = 1 - \mathrm{e}^{2(\lambda - \lambda_0)} \tag{A.11}$$

where we have introduced the maximum Lyapunov exponent $\lambda_0 = \log g\sigma$. We can finally invert this expression obtaining

$$\lambda = \lambda_0 + \frac{1}{2}\log(1 - \rho) \,. \tag{A.12}$$

This formula gives the shape of the Lyapunov spectrum, under the assumption that the eigenvalues are distributed according to Girko's law. The Lyapunov spectrum decreases linearly in the region around the maximum. Moreover, the shape is basically parameter-free, except for a vertical shift that is related to the maximum exponent.

In the more general case $a > 0$, q_e is larger than zero already in the microscopic fixed point, so that the hyperbolic cosine comes into play. In a first approximation, one

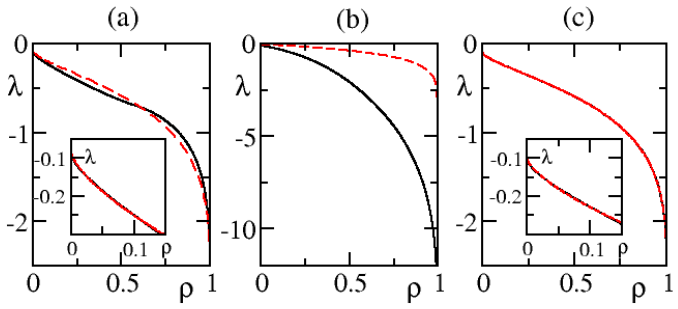

Figure A.3: (color online) Lyapunov spectrum of the stable fixed point in cases (I), (II) and (III) (panels, a, b, and c, respectively - solid lines). The three spectra have been obtained for $g = 0.6$, 2.1 and 0.4, respectively (all averaged over 100 different realizations and for $N = 1024$). The dashed lines correspond to suitable theoretical approximation (see the text).

can assume that the local fields are statistically independent, as well as independent of the J_{ij} elements and thereby invoke Girko's theorem (this was done in [57] for case (II)). As a result, the radius of the distribution coincides with the standard deviation σ of the matrix elements, i.e., by invoking the mean-field approximation,

$$\sigma^2 = \left\langle (\cosh q_e h)^{-4} \right\rangle \ , \tag{A.13}$$

where the value of q_e is determined self-consistently from Eq. (A.5), so that the stability condition becomes $\lambda_0 = \log(\gamma + \sigma g_c) > 0$.

The spectra plotted in Fig. A.3a,b (dashed curves) reveal that this is an excellent approximation of the maximum exponent, but much less so for the overall shape. In fact, a direct inspection of the eigenvalues of $\mathbf{M}$, reveals that for $a > 0$ they are still characterized by a radially symmetric distribution, but their density is no longer uniform within the disk: it increases upon approaching the center of the disk, where it even exhibits a singularity, when $a = 1$ (see Fig. A.4). Anyway, since there is no significant density variation close to border of the disk, the Lyapunov spectrum keeps a linear shape around the maximum as in the previous case.

For $\gamma > 0$, the eigenvalues are still distributed within a disk, but its center is shifted to $(\gamma, 0)$ (see Fig. A.5 for the case $\sigma = 1$ and $\gamma = 0.5$). Accordingly, the eigenvalues along a ring of given radius do not have the same modulus, and the relationship with the Lyapunov exponent becomes more complicated. Therefore, instead of deriving theoretical expressions, we have carried out the calculations in a purely numerical way. The result is the dashed line plotted in Fig. A.3c, where it is hardly distinguishable from the numerical observations. This confirms the validity of the theoretical analysis. Since the largest moduli are now located in a smaller region (see the shaded area in Fig. A.5) the initial part of the Lyapunov spectrum decreases in a faster way. It is easy to see that the number of exponents, whose modulus differ less than $\delta\mu$ from the maximum grows as $\delta\mu^{3/2}$. Upon exchanging the independent with the dependent variable, we find that $\lambda = \lambda_0 - \alpha\rho^{2/3}$, as confirmed in Fig. A.3,

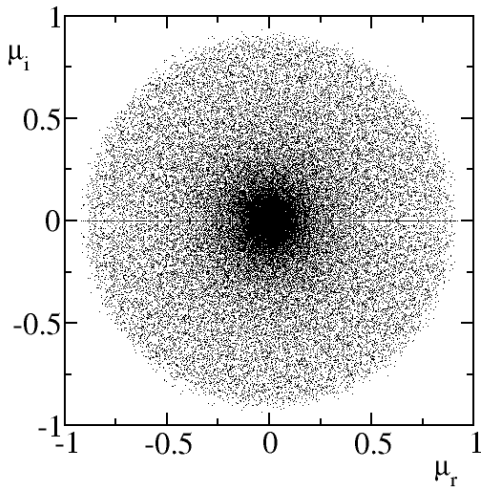

Figure A.4: Distribution of the eigenvalues for the same parameter values as in panel b of Fig. A.3.

where this functional dependence reproduces the initial part of the spectrum (see the two insets) with an excellent agreement.

A.5 The chaotic phase

Above the critical value g_c, the fixed point becomes unstable and a time-dependent state sets in. The best way to characterize this regime is in terms of the Lyapunov exponents. We start from the maximum one. In Fig. A.6a we see that λ_0 depends weakly on the system size and eventually converges towards an asymptotic curve that remains below the dashed line, which corresponds to the maximum exponent of the fixed point. Below g_c, the theoretical prediction is very close to the numerical results: this confirms indirectly the assumed statistical independence among the local fields. The fact that, above criticality, the instability of the fixed point is larger than the actual maximum Lyapunov exponent, is a reasonable expectation: it indicates that generic trajectories escape from the unstable fixed point to settle in a less unstable region.

The results reported in Fig. A.6b are definitely more puzzling, as the disagreement below criticality suggests that the conjectured statistical-independence might be wrong. Strange enough, a much better agreement is observed above criticality, where, a priori, there is no reason to expect it. Altogether, it appears that the mean-field approach provides a rather accurate characterization of the chaotic dynamics. From the point of view of the critical behavior, we see that in both cases, λ_0 increases linearly with $g - g_c$, suggesting that the shape of the spectrum around the maximum (see Fig. A.3) does not affect the scaling behaviour of the maximum exponent (we further comment on this point in the following).

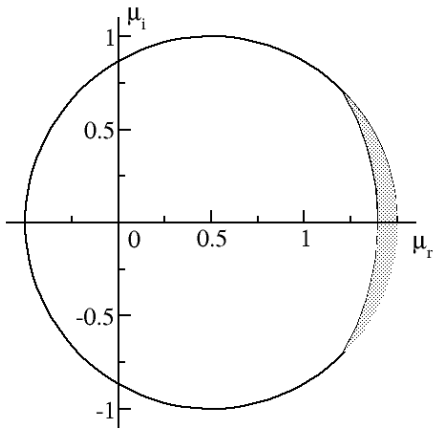

Figure A.5: Disk with the eigenvalues in the case $\gamma = 0.5$ and $\sigma = 1$. The shaded area correspond to the eigenvalues with a modulus larger than a given value close to the maximum.

Panels c and d of Fig. A.6 refer to the cases (III) and (IV), where $a = 0$. They both reveal a quadratic growth of λ_0 on $g - g_c$. A possible motivation for the different behavior is that, at variance with the previous cases, here the mean field equation reveals a macroscopic transition (the ensemble variance starts to grow from zero above the critical point). A perturbative calculation of the maximum Lyapunov exponent of the fixed point reveals that $\lambda_0 = (g - g_c)^2/2$, as also predicted in [232] for case (IV). The direct comparison between the theoretical prediction (see the dashed curves) and numerical observations show, nevertheless, the presence of increasingly relevant deviations, when the coupling strength is increased above the critical value. Most of the discrepancy is presumably to be attributed to stronger finite-size corrections. Notice, finally, that the λ_0 values are quite small in panel *d*. This is because in the continuous-time limit, it is more appropriate to rescale the time by a factor $1/(1-\gamma)$. Altogether, the study of λ_0 confirms the prediction of the transition point estimated from the onset of a finite temporal variance q_t and reveals that the dynamics becomes immediately chaotic above the critical point.

More detailed information on the chaotic behaviour can be obtained by computing the entire Lyapunov spectrum. The need to average over a certain number (typically about 100) of realizations of the network structures makes this task computationally very demanding. Therefore, we have proceeded by performing the simulations on a Nvidia GTX680 GPU equipped with a Kepler architecture [73]. Computer simulations have been performed in CUDA-C language [211, 40]. Preliminary studies have shown that the performance depends crucially on the appropriate use of the NVIDIA CUDA BLAS routines [73]. In order to optimize the determination of the Lyapunov exponents, we have made use of a QR decomposition available for GPU [48]. The Ly-

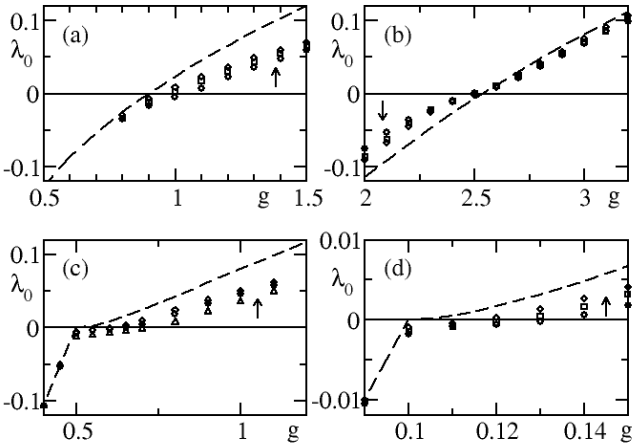

Figure A.6: Maximum Lyapunov exponent in the four cases (I-IV), with the same notations as in the first figure. The dashed line is the result of a mean field approach (see the text).

punov spectra for the cases (I-III) are reported in Fig. A.7, each panel corresponding to a single coupling strength, above the transition (for details see the figure caption). The different curves in each panel correspond to different system sizes: the nice overlap reveals that the chaotic dynamics is extensive, i.e. the fractal dimension as well as the Kolmogorov-Sinai entropy are both proportional to the network size. The insets provide a clean evidence of the convergence towards a smooth shape (at least for the initial part of the spectrum). Extensivity is a standard property of spatio-temporal chaos in systems with short-range coupling, where it follows naturally from the natural idea that the dynamics of a full chain can be broken into that of almost independent subchains. Extensivity is much less obvious in non-additive sytems, such as the present setup or in networks with a sparse connectivity (see, e.g., [173]): a clean explanation in this latter context is still missing.

The very fact that the initial slope of the spectrum is finite indicates that the appearance of active degrees of freedom upon increasing the coupling strength can be understood by just looking at the maximum Lyapunov exponent λ_0. As long as λ_0 increases linearly from zero (cases I-II), so it does the density of dimension, while the entropy increases quadratically. This is exactly what found in case (II) (see Table A.1).

The critical behavior in the setups III-IV is less clear, since below the transition, the Lyapunov spectrum exhibits a singular behaviour (with the above mentioned exponent 2/3), so that one should analyse the cross-over to the linear shape, before making a convincing statement. This, combined with the weaker growth (quadratic) of the chaotic component, makes the numerical analysis rather problematic and we, in fact, do not dare to present any conjecture.

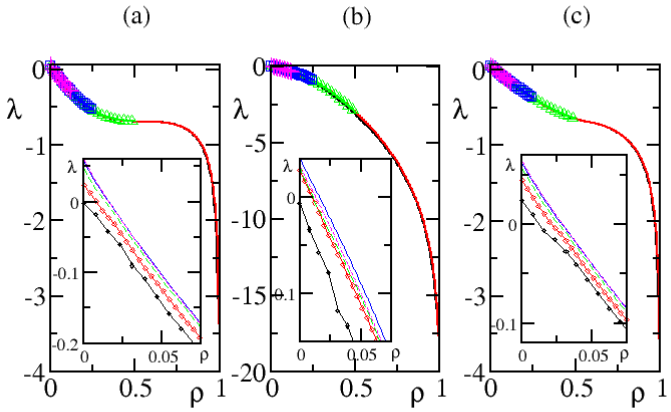

Figure A.7: (color online) Lyapunov spectra for the cases (I-III). Panels (a-c) correspond to $g = 1.4, 2.75, 1.1$, respectively. In all main panels the different colors correspond to different network sizes: black, red, and green triangles, blue squares and magenta plusses correspond respectively to $N = 128, 256, 512, 1024, 2048$. In the insets, black circles, red diamonds, green dashed, straight blue and dashed-dot magenta lines, correspond to $N = 128, 256, 512, 1024, 2048$. All the spectra are the average over 64 different realizations.

Table A.1: Density of dimension and entropy in case (II)

gauge g	d_{KY}	h_{KS}
2.65	0.015	0.0013
2.75	0.026	0.0044
3.0	0.05	0.02
3.25	0.079	0.046

Finally, we have looked directly at the dynamics of the local fields by studying their correlation function. In fact, the relationship between the number of degrees of freedom involved in the evolution of a given observable and its correlation properties is far from obvious: there may be high-dimensional systems accompanied by power-law decays of the correlations (see e.g. heat transport in one dimension [172]), or one-dimensional maps with δ-like correlations (see, e.g., the symmetric tent map [199]). Let us start defining the rescaled correlation,

$$C(\tau) = \frac{\langle h_i(t) h_i(t+\tau)\rangle_t - \langle h_i(t)\rangle_t^2}{\langle h_i^2(t)\rangle_t - \langle h_i(t)\rangle_t^2}\,, \tag{A.14}$$

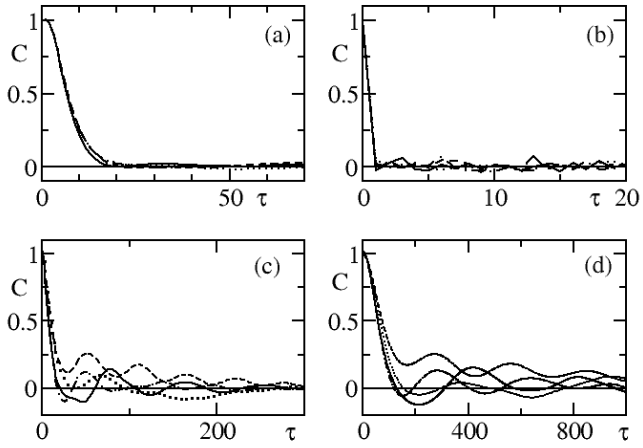

Figure A.8: Normalized autocorrelations for the 4 cases (I-IV). Panels (a-d) correspond to $g = 1.4$, 3., 0.8 and 0.15, respectively. All simulations have been performed in a lattice with 2048 neurons; the four curves correspond to 4 different realizations of the network.

where we have assumed that all neurons are characterized (in the thermodynamic limit) by the same correlation. This is one of the key points of the theoretical analysis performed in [232] for case (IV), where a self-consistent equation for the full $C(\tau)$ has been derived. The independence of $C(\tau)$ on the index i in cases (I) and (II) is less obvious, since the single neurons are characterized by different non-zero steady local fields. Our numerical simulations, however, suggest that the shape of the rescaled correlation function is common to all neurons. This can be appreciated in Fig. A.8, where four random examples are plotted for each of the four different cases. More than that, the fluctuations among the different neurons are smaller in panels (a) and (b), which refer to setups where the single local field differ from each other. Furthermore, we have verified that the decay in case (IV) is consistent with the theoretical expectations of Ref. [232] and the same applies to case (III). Finally, the very fast decay (almost one time unit) observed in case (II) is consistent with the fact that the instability there arises from the simultaneous instability of many different frequencies, since for $a = 0$ the unit disk is centered in the origin.

A.6 Conclusions and open problems

In this paper we have performed a detailed analysis of a neuronal model, where the local mean field has not a preferred sign (i.e., it is neither strictly excitatory nor inhibitory). Such a regime is often referred to as a balanced state. Our model combines into a single framework two different systems previously studied in the literature: it reduces to them for suitable choices of the parameter values.

With the help of a mean-field analysis, borrowing ideas from the spectra of random matrices and with the help of direct numerical simulations we find that two qualitatively different scenarios are observed, depending whether an effective external random field (gauged by the constant a) is present or not: in the former case, the existence of a phase transition can be inferred by monitoring the *temporal* variance, while the *ensemble* variance increases smoothly with the coupling strength. A second difference between the two regimes is given by the scaling behavior of the maximum Lyapunov exponent, which increases quadratically (with the distance from the critical point) when $a = 0$, while it increases linearly in the more general case $a > 0$. Moreover, even the characterization of the stable regime requires further studies. The theoretical predictions based on the random matrix theory are quite accurate for what concerns the maximum Lyapunov exponents, but it is not clear whether the theoretical prediction is exact. We cannot exclude that subtle correlations among the local mean field might invalidate the application of the formula. As for the full spectrum of exponents, it is instead definitely clear that it is necessary to go beyond the standard circle's law. The numerically observed radial symmetry suggests the possibility to develop some clever conjecture.

A second crucial parameter is the local contraction γ. On the one hand, for $\gamma > 0$, the Lyapunov spectrum decreases as $\rho^{-2/3}$ away from the maximum (below criticality), while it decreases linearly when $\gamma = 0$. On the other hand, we numerically observe that above criticality, the Lyapunov spectrum has a linear shape in both cases. The way the power-law behaviour crosses over to a linear behavior for $\gamma = 0$ is still an open point that requires further more refined studies.

Another interesting direction to explore is the extension of the dynamical mean-field developed in Ref. [232] to cover the case where the time variable is discrete and, more difficult to achieve, to account for the presence of an external random field, which induces a substantial change in the transition scenario.

Furthermore, it would be desirable to identify the effective degrees of freedom that are switched on, when the collective motion sets in: our simulations indicate that the number of active degrees of freedom is proportional to the network size, but it is not clear how they are related to the network structure. Some light on this problem is given by the correlation function of the local field, which indicates that a strong dissipation is accompanied also by a fast decay and one might thus develop some more refined theoretical arguments. In any case, a deeper understanding of the chaotic phase is far from trivial, as suggested by the shape of the Lyapunov spectrum which changes substantially when the coupling strength is increased above the transition.

Finally, it would be nice to repeat this analysis in more microscopic models such as networks of pulse coupled oscillators, in order to test the generality of our observations. A preliminary study in such a direction can be found in [189], but a characterization close to criticality is definitely an ambitious task.

Bibliography

[1] Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the joint Advisory Committee on Emerging Regulatory Issues. Tech. rep., Commodity Futures Trading Commission and Securities and Exchange Commission, September 30 2010.

[2] *Global Optimization Toolbox User's Guide*. MathWorks, 2012.

[3] *Optimization Toolbox User's Guide*. MathWorks, 2012.

[4] *Parallel Computing Toolbox User's Guide*. MathWorks, 2012.

[5] Report to Congress on Decimalization. Tech. rep., U.S. Securities and Exchange Commission, July 2012.

[6] Abbasbandy, S. The application of homotopy analysis method to solve a generalized Hirota-Satsuma coupled KdV equation. *Physics Letters A 361*, 6 (2007), 478–483.

[7] Abbasbandy, S. Homotopy analysis method for heat radiation equations. *International Communications in Heat and Mass Transfer 34*, 3 (2007), 380–387.

[8] Abbasbandy, S., and Shivanian, E. A new analytical technique to solve Fredholm's integral equations. *Numerical algorithms 56*, 1 (2011), 27–43.

[9] Abbasbandy, S., Shivanian, E., and Vajravelu, K. Mathematical properties of $\hbar$-curve in the frame work of the homotopy analysis method. *Communications in Nonlinear Science and Numerical Simulation 16*, 11 (2011), 4268 – 4275.

[10] Abergel, F., Bouchaud, J.-P., Foucault, T., Lehalle, C.-A., and Rosenbaum, M. *Market microstructure: confronting many viewpoints*. Wiley, 2012.

[11] Adomian, G. A review of the decomposition method and some recent results for nonlinear equations. *Mathematical and Computer Modelling 13*, 7 (1990), 17–43.

[12] Ahn, H.-J., Cai, J., Chan, K., and Hamao, Y. Tick size change and liquidity provision on the Tokyo stock exchange. *Journal of the Japanese and International Economies 21*, 2 (2007), 173 – 194.

[13] Akyildiz, F. T., and Vajravelu, K. Magnetohydrodynamic flow of a viscoelastic fluid. *Physics Letters A 372*, 19 (2008), 3380 – 3384.

[14] Alfonsi, A., Schied, A., and Slynko, A. Order book resilience, price manipulation, and the positive portfolio problem. *SIAM Journal on Financial Mathematics 3*, 1 (2012), 511–533.

[15] Almgren, R., and Chriss, N. Value under liquidation. *Risk* (December 1999), 61–63.

[16] Almgren, R., and Chriss, N. Optimal Execution of Portfolio Transactions. *The Journal of Risk* (2000), 5–39.

[17] Almgren, R., Thum, C., Hauptmann, E., and Li, H. Direct estimation of equity market impact. *Risk 18* (2005), 5752.

[18] Almgren, R. F. Optimal execution with nonlinear impact functions and trading-enhanced risk. *Applied Mathematical Finance 10*, 1 (2003), 1–18.

[19] Andersen, T. G., and Bollerslev, T. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance 4*, 23 (1997), 115 – 158. High Frequency Data in Finance, Part 1.

[20] Ané, T., and Geman, H. Order flow, transaction clock, and normality of asset returns. *The Journal of Finance 55*, 5 (2000), 2259–2284.

[21] Antoniou, A., and Vorlow, C. E. Price clustering and discreteness: is there chaos behind the noise? *Physica A: Statistical Mechanics and its Applications 348* (2005), 389–403.

[22] Arenas, A., Díaz-Guilera, A., Jurgen, K., Yamir, M., and Changsong, Z. Synchronization in complex networks. *Physics Reports 469*, 3 (2008), 93 – 153.

[23] Ascioglu, A., Comerton-Forde, C., and McInish, T. H. An examination of minimum tick sizes on the Tokyo stock exchange. *Japan and the World Economy 22*, 1 (2010), 40 – 48.

[24] Awrejcewicz, J., Andrianov, I. V., and Manevitch, L. I. *Asymptotic Approaches in Nonlinear Dynamics*. Springer-Verlag Berlin Heidelberg, 1998.

[25] Bachelier, L. Théorie de la spéculation. *Ann. Sci. Éc. Norm. Supér. (3) 17* (1900), 21–86.

[26] Bacry, E., Iuga, A., Lasnier, M., and Lehalle, C.-A. Market impacts and the life cycle of investors orders. *Available at SSRN 2532152* (Nov. 2014).

[27] Baillie, R. T. Long memory processes and fractional integration in econometrics. *Journal of econometrics 73*, 1 (1996), 5–59.

[28] Basharin, G. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications 4*, 3 (1959), 333–336.

[29] Basseville, M. Divergence measures for statistical data processing - An annotated bibliography. *Signal Processing 93*, 4 (2013), 621 – 633.

[30] Berchtold, A. The double chain Markov model. *Communications in Statistics - Theory and Methods 28*, 11 (1999), 2569–2589.

[31] Berchtold, A. High-order extensions of the double chain Markov model. *Stochastic Models 18*, 2 (2002), 193–227.

[32] Berchtold, A., and Raftery, A. The mixture transition distribution model for high-order Markov chains and non-gaussian time series. *Statist. Sci. 17*, 3 (08 2002), 328–356.

[33] Berger, M. S. *Nonlinearity & Functional Analysis: Lectures on Nonlinear Problems in Mathematical Analysis*, vol. 74. Academic press, 1977.

[34] Bershova, N., and Rakhlin, D. The non-linear market impact of large trades: evidence from buy-side order flow. *Quantitative Finance 13*, 11 (2013), 1759–1778.

[35] Bertsimas, D., and Lo, A. W. Optimal control of execution costs. *Journal of Financial Markets 1*, 1 (1998), 1 – 50.

[36] Bessembinder, H. The degree of price resolution and equity trading costs. *Journal of Financial Economics 45*, 1 (1997), 9 – 34. Symposium on Market Microstructure: Focus on Nasdaq.

[37] Bessembinder, H. Trade execution costs on NASDAQ and the NYSE: A post-reform comparison. *Journal of Financial and Quantitative Analysis 34* (9 1999), 387–407.

[38] Bessembinder, H. Tick size, spreads, and liquidity: An analysis of Nasdaq securities trading near ten dollars. *Journal of Financial Intermediation 9*, 3 (2000), 213 – 239.

[39] Bessembinder, H. Trade execution costs and market quality after decimalization. *Journal of Financial and Quantitative Analysis 38* (12 2003), 747–777.

[40] Block, B., and Preis, T. Computer simulations of the Ising model on graphics processing units. *The European Physical Journal-Special Topics 210*, 1 (2012), 133–145.

[41] Bollen, N. P. B., and Busse, J. A. Tick size and institutional trading costs: Evidence from mutual funds. *Journal of Financial and Quantitative Analysis 41* (12 2006), 915–937.

[42] Bouchaud, J.-P., Farmer, J. D., and Lillo, F. How markets slowly digest changes in supply and demand. *Handbook of financial markets: dynamics and evolution 1* (2009), 57.

[43] Bouchaud, J.-P., Gefen, Y., Potters, M., and Wyart, M. Fluctuations and response in financial markets: the subtle nature of random price changes. *Quantitative Finance 4*, 2 (2004), 176–190.

[44] Bouchaud, J.-P., Kockelkoren, J., and Potters, M. Random walks, liquidity molasses and critical response in financial markets. *Quantitative Finance 6*, 2 (2006), 115–123.

[45] Bouchaud, J.-P., Mézard, M., and Potters, M. Statistical properties of stock order books: empirical results and models. *Quantitative Finance 2*, 4 (2002), 251–256.

[46] Bouchaud, J.-P., and Potters, M. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge University Press, 2003.

[47] Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. *Time Series Analysis: Forecasting and Control. 3rd ed.* Prentice Hall, 1994.

[48] Brandes, T., Arnold, A., Soddemann, T., and Reith, D. Cpu vs. gpu-performance comparison for the gram-schmidt algorithm. *The European Physical Journal Special Topics 210*, 1 (2012), 73–88.

[49] Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences 106*, 30 (2009), 12267–12272.

[50] Brokmann, X., Kockelkoren, J., and Bouchaud, J.-P. Slow decay of impact in equity markets. *Available at SSRN 2471528* (2014).

[51] Browder, F., and Petryshyn, W. Construction of fixed points of nonlinear mappings in hilbert space. *Journal of Mathematical Analysis and Applications 20*, 2 (1967), 197–228.

[52] Brown, K., and Sethna, J. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E 68* (Aug 2003), 021904.

[53] Bulla, J., and Bulla, I. Stylized facts of financial time series and hidden semi-markov models. *Computational Statistics & Data Analysis 51*, 4 (2006), 2192 – 2209. Nonlinear Modelling and Financial Econometrics.

[54] Busseti, E., and Lillo, F. Calibration of optimal execution of financial transactions in the presence of transient market impact. *Journal of Statistical Mechanics: Theory and Experiment 2012*, 09 (2012), P09010.

[55] Cappé, O., Moulines, E., and Rydén, T. *Inference in hidden Markov models*, vol. 6. Springer, 2005.

[56] Carpi, L. C., Rosso, O. A., Saco, P. M., and Ravetti, M. G. Analyzing complex networks evolution through information theory quantifiers. *Physics Letters A 375*, 4 (2011), 801 – 804.

[57] Cessac, B., Doyon, B., Quoy, M., and Samuelides, M. Mean-field equations, bifurcation map and route to chaos in discrete time neural networks. *Physica D: Nonlinear Phenomena 74*, 12 (1994), 24 – 44.

[58] Chan, K. S., and Tong, H. On estimating thresholds in autoregressive models. *Journal of time series analysis 7*, 3 (1986), 179–190.

[59] Chekmarev, S. Information entropy as a measure of nonexponentiality of waiting-time distributions. *Phys. Rev. E 78* (Dec 2008), 066113.

[60] Chiang, A. C. *Fundamental methods of mathematical economics*. McGraw-Hill, 1984.

[61] Christie, W. G., and Schultz, P. H. Why do NASDAQ market makers avoid odd-eighth quotes? *The Journal of Finance 49*, 5 (1994), 1813–1840.

[62] Chung, K. H., Chuwonganant, C., and McCormick, D. Order preferencing and market quality on NASDAQ before and after decimalization. *Journal of Financial Economics 71*, 3 (2004), 581 – 612.

[63] Clark, P. K. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica: journal of the Econometric Society* (1973), 135–155.

[64] Clauset, A., Shalizi, C., and Newman, M. Power-law distributions in empirical data. *SIAM Review 51*, 4 (2009), 661–703.

[65] Clauset, A., Young, M., and Gleditsch, K. S. On the frequency of severe terrorist events. *Journal of Conflict Resolution 51*, 1 (2007), 58–87.

[66] Cogburn, R. The ergodic theory of Markov chains in random environments. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 66*, 1 (1984), 109–128.

[67] Cogburn, R. On direct convergence and periodicity for transition probabilities of Markov chains in random environments. *The Annals of Probability* (1990), 642–654.

[68] Connor, R., Cardillo, F. A., Moss, R., and Rabitti, F. Evaluation of Jensen-Shannon distance over sparse data. In *Similarity Search and Applications*. Springer, 2013, pp. 163–168.

[69] Cont, R. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance 1*, 2 (2001), 223–236.

[70] Cont, R., and De Larrard, A. Price dynamics in a markovian limit order market. *SIAM Journal on Financial Mathematics 4*, 1 (2013), 1–25.

[71] Cont, R., Stoikov, S., and Talreja, R. A stochastic model for order book dynamics. *Operations research 58*, 3 (2010), 549–563.

[72] Contreras-Reyes, J. E., and Arellano-Valle, R. B. Kullback–Leibler divergence measure for multivariate skew-normal distributions. *Entropy 14*, 9 (2012), 1606–1626.

[73] Cook, S. *CUDA programming: a developer's guide to parallel computing with GPUs*. Newnes, 2012.

[74] Cover, T. M., and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

[75] Cristelli, M., Alfi, V., P., and Zaccaria, A. Liquidity crisis, granularity of the order book and price fluctuations. *Eur. Phys. J. B 73*, 1 (2010), 41–49.

[76] Crooks, G. E. Measuring thermodynamic length. *Physical review letters 99*, 10 (2007), 100602.

[77] Crooks, G. E., and Sivak, D. A. Measures of trajectory ensemble disparity in nonequilibrium statistical dynamics. *Journal of Statistical Mechanics: Theory and Experiment 2011*, 06 (2011), P06003.

[78] D'Amico, G., and Petroni, F. A semi-Markov model with memory for price changes. *Journal of Statistical Mechanics: Theory and Experiment 2011*, 12 (2011), P12009.

[79] D'Amico, G., and Petroni, F. A semi-Markov model for price returns. *Physica A: Statistical Mechanics and its Applications 391*, 20 (2012), 4867 – 4876.

[80] D'Amico, G., and Petroni, F. Weighted-indexed semi-Markov models for modeling financial returns. *Journal of Statistical Mechanics: Theory and Experiment 2012*, 07 (2012), P07015.

[81] Dang, N. M. Optimal Execution with Transient Impact. `http://ssrn.com/abstract=2183685`, 2014.

[82] Daniels, M., Farmer, J., Gillemot, L., Iori, G., and Smith, E. Quantitative model of price diffusion and market friction based on trading as a mechanistic random process. *Phys. Rev. Lett. 90* (Mar 2003), 108102.

[83] Dayri, K., Bacry, E., and Muzy, J.-F. *The Nature of Price Returns During Periods og High Market Activity*. Econophysics of Order-driven Markets. Springer, 2011, p. 155.

[84] Dayri, K., and Rosenbaum, M. Large tick assets: implicit spread and optimal tick size. *arXiv preprint arXiv:1207.6325* (2012).

[85] De Domenico, M., and Insolia, A. Entropic approach to multiscale clustering analysis. *Entropy 14*, 5 (2012), 865–879.

[86] Deco, G., and Hugues, E. Balanced input allows optimal encoding in a stochastic binary neural network model: An analytical study. *PLoS ONE 7*, 2 (02 2012), e30723.

[87] Donier, J., Bonart, J., Mastromatteo, I., and Bouchaud, J.-P. A fully consistent, minimal model for non-linear market impact. *arXiv:1412.0141v4* (2015).

[88] Easley, D., and O'Hara, M. Time and the process of security price adjustment. *The Journal of finance 47*, 2 (1992), 577–605.

[89] Eisenkopf, A. The Real Nature of Credit Rating Transitions. `http://ssrn.com/abstract=968311`, 2008.

[90] Eisler, Z., Bouchaud, J.-P., and Kockelkoren, J. The price impact of order book events: market orders, limit orders and cancellations. *Quantitative Finance 12*, 9 (2012), 1395–1419.

[91] Eisler, Z., Perelló, J., and Masoliver, J. Volatility: A hidden Markov process in financial time series. *Phys. Rev. E 76* (Nov 2007), 056105.

[92] Endres, D. M., and Schindelin, J. E. A new metric for probability distributions. *Information Theory, IEEE Transactions on 49*, 7 (2003), 1858–1860.

[93] Engle, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*, 4 (1982), 987–1007.

[94] Engle, R. F., and Russell, J. R. Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica* (1998), 1127–1162.

[95] Farmer, D. J., Gillemot, L., Lillo, F., Mike, S., and Sen, A. What really causes large price changes? *Quantitative Finance 4*, 4 (2004), 383–397.

[96] Farmer, J. D., Gerig, A., Lillo, F., and Mike, S. Market efficiency and the long-memory of supply and demand: is price impact variable and permanent or fixed and temporary? *Quantitative Finance 6*, 2 (2006), 107–112.

[97] Farmer, J. D., Gerig, A., Lillo, F., and Waelbroeck, H. How efficiency shapes market impact. *Quantitative Finance 13*, 11 (2013), 1743–1758.

[98] Felizzi, F., and Comoglio, F. Network-of-queues approach to b-cell-receptor affinity discrimination. *Phys. Rev. E 85* (Jun 2012), 061926.

[99] Ferland, R., Latour, A., and Oraichi, D. Integer-valued GARCH process. *Journal of Time Series Analysis 27*, 6 (2006), 923–942.

[100] Fitzpatrick, M., and Marchev, D. Efficient bayesian estimation of the multivariate double chain Markov model. *Statistics and Computing 23*, 4 (2013), 467–480.

[101] Forsyth, P. A., Kennedy, J. S., Tse, S., and Windcliff, H. Optimal trade execution: a mean quadratic variation approach. *Journal of Economic Dynamics and Control 36*, 12 (2012), 1971–1991.

[102] Foucault, T., Kadan, O., and Kandel, E. Limit order book as a market for liquidity. *Review of Financial Studies 18*, 4 (2005), 1171–1217.

[103] Frühwirth-Schnatter, S. *Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes*. Springer, 2006.

[104] Gatheral, J. No-dynamic-arbitrage and market impact. *Quantitative Finance 10*, 7 (2010), 749–759.

[105] Gatheral, J., and Schied, A. Optimal trade execution under geometric brownian motion in the Almgren and Chriss framework. *International Journal of Theoretical and Applied Finance 14*, 03 (2011), 353–368.

[106] Gatheral, J., and Schied, A. Dynamical models of market impact and algorithms for order execution. In *Handbook on Systemic Risk*, J.-P. Fouque and J. A. Langsam, Eds. Cambridge University Press, 2013, pp. 579–602. Cambridge Books Online.

[107] Gatheral, J., Schied, A., and Slynko, A. Exponential resilience and decay of market impact. In *Econophysics of order-driven markets*. Springer, 2011, pp. 225–236.

[108] Gatheral, J., Schied, A., and Slynko, A. Transient linear price impact and Fredholm integral equations. *Mathematical Finance 22*, 3 (2012), 445–474.

[109] Gençay, R., Dacorogna, M., Muller, U. A., Pictet, O., and Olsen, R. *An introduction to high-frequency finance*. Academic Press, 2001.

[110] Gerig, A. A theory for market impact: How order flow affects stock price. *arXiv preprint arXiv:0804.3818* (2008).

[111] Geweke, J., and Porter-Hudak, S. The estimation and application of long memory time series models. *Journal of Time Series Analysis 4*, 4 (1983), 221–238.

[112] Gibson, S., Singh, R., and Yerramilli, V. The effect of decimalization on the components of the bid-ask spread. *Journal of Financial Intermediation 12*, 2 (2003), 121–148.

[113] Gillemot, L., Farmer, J. D., and Lillo, F. There's more to volatility than volume. *Quantitative Finance 6*, 5 (2006), 371–384.

[114] Girko, V. L. Circular law. *Teoriya Veroyatnostei I Ee Primeneniya 29*, 4 (1984), 669–679.

[115] Gleser, L. J. Exact power of goodness-of-fit tests of Kolmogorov type for discontinuous distributions. *Journal of the American Statistical Association 80*, 392 (1985), 954–958.

[116] Glosten, L. R., and Milgrom, P. R. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics 14*, 1 (1985), 71 – 100.

[117] Goldstein, M. A., and Kavajecz, K. A. Eighths, sixteenths, and market depth: changes in tick size and liquidity provision on the NYSE. *Journal of Financial Economics 56*, 1 (2000), 125 – 149.

[118] Gopikrishnan, P., Plerou, V., Nunes, A. L., Meyer, M., and Stanley, H. Scaling of the distribution of fluctuations of financial market indices. *Phys. Rev. E 60* (Nov 1999), 5305–5316.

[119] Granger, C. W. J., and Andersen, A. P. *An introduction to bilinear time series models*. Vandenhoeck & Ruprecht, 1978.

[120] Granger, C. W. J., and Orr, D. Infinite variance and research strategy in time series analysis. *Journal of the American Statistical Association 67*, 338 (1972), 275–285.

[121] Granger, C. W. J., and Terasvirta, T. Modelling non-linear economic relationships. *OUP Catalogue* (1993).

[122] Grassberger, P. Finite sample corrections to entropy and dimension estimates. *Physics Letters A 128*, 67 (1988), 369 – 373.

[123] Gross-KlussMann, A., and Hautsch, N. Predicting bid-ask spreads using long-memory autoregressive conditional Poisson models. *Journal of Forecasting 32*, 8 (2013), 724–742.

[124] Grosse, I., Bernaola-Galván, P., Carpena, P., Román-Roldán, R., Oliver, J., and Stanley, H. Analysis of symbolic sequences using the Jensen-Shannon divergence. *Phys. Rev. E 65* (Mar 2002), 041905.

[125] Guilbaud, F., and Pham, H. Optimal high-frequency trading with limit and market orders. *Quantitative Finance 13*, 1 (2013), 79–94.

[126] Hall, A. D., and Hautsch, N. Order aggressiveness and order book dynamics. *Empirical Economics 30*, 4 (2006), 973–1005.

[127] Hamilton, J. D. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society* (1989), 357–384.

[128] Hamilton, J. D. *Time series analysis*, vol. 2. Princeton University Press, Princeton, 1994.

[129] Harris, L. Stock price clustering and discreteness. *Review of Financial Studies 4*, 3 (1991), 389–415.

[130] Harris, L. *Trading and exchanges: Market microstructure for practitioners*. Oxford University Press, 2002.

[131] Harris, L., and Hasbrouck, J. Market vs. limit orders: The superdot evidence on order submission strategy. *Journal of Financial and Quantitative Analysis 31* (6 1996), 213–231.

[132] Hasbrouck, J. Measuring the information content of stock trades. *The Journal of Finance 46*, 1 (1991), 179–207.

[133] Hasbrouck, J. The dynamics of discrete bid and ask quotes. *The Journal of finance 54*, 6 (1999), 2109–2142.

[134] Hasbrouck, J. *Empirical market microstructure: The institutions, economics, and econometrics of securities trading*. Oxford University Press, 2006.

[135] Hautsch, N. *Econometrics of financial high-frequency data*. Springer, 2011.

[136] Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika 58*, 1 (1971), 83–90.

[137] He, Y., and Wu, C. Price rounding and bid–ask spreads before and after the decimalization. *International Review of Economics & Finance 13*, 1 (2004), 19–41.

[138] Herzel, H., Schmitt, A., and Ebeling, W. Finite sample effects in sequence analysis. *Chaos, Solitons & Fractals 4*, 1 (1994), 97–113.

[139] Hetmaniok, E., Słota, D., Trawiński, T., and Wituła, R. Usage of the homotopy analysis method for solving the nonlinear and linear integral equations of the second kind. *Numerical Algorithms* (2013), 1–23.

[140] Hill, B. M. A simple general approach to inference about the tail of a distribution. *The annals of statistics 3*, 5 (1975), 1163–1174.

[141] House, M., Jiang, H., and Zhang, X. Analysis of electron tunneling events with the hidden Markov model. *Phys. Rev. B 80* (Sep 2009), 113308.

[142] Huang, R. D., and Stoll, H. R. Tick size, bid-ask spreads, and market structure. *Journal of Financial and Quantitative Analysis 36*, 04 (2001), 503–522.

[143] Huberman, G., and Stanzl, W. Price manipulation and quasi-arbitrage. *Econometrica 72*, 4 (2004), 1247–1275.

[144] Jahnke, S., Memmesheimer, R.-M., and Timme, M. Stable irregular dynamics in complex neural networks. *Phys. Rev. Lett. 100* (Jan 2008), 048102.

[145] Johnson, N. L., Kemp, A. W., and Kotz, S. *Univariate discrete distributions*, vol. 444. John Wiley & Sons, 2005.

[146] Kedem, B., and Fokianos, K. *Regression models for time series analysis*, vol. 488. John Wiley & Sons, 2005.

[147] Kempf, A., and Korn, O. Market depth and order size. *Journal of Financial Markets 2*, 1 (1999), 29 – 48.

[148] Klöck, F., Schied, A., and Sun, Y. Price manipulation in a market impact model with dark pool. *Available at SSRN 1785409* (2011).

[149] Kolda, T., Lewis, R., and Torczon, V. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review 45*, 3 (2003), 385–482.

[150] Kolda, T., Lewis, R., and Torczon, V. Stationarity results for generating set search for linearly constrained optimization. *SIAM Journal on Optimization 17*, 4 (2007), 943–968.

[151] Korobkova, E., Emonet, T., Park, H., and Cluzel, P. Hidden stochastic nature of a single bacterial motor. *Phys. Rev. Lett. 96* (Feb 2006), 058105.

[152] Kullback, S. *Information theory and statistics*. Courier Dover Publications, 1997.

[153] Kyle, A. S. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society* (1985), 1315–1335.

[154] La Spada, G., Farmer, J. D., and Lillo, F. Tick size and price diffusion. In *Econophysics of Order-driven Markets*. Springer, 2011, pp. 173–187.

[155] La Spada, G., and Lillo, F. The effect of round-off error on long memory processes. *Studies in Nonlinear Dynamics and Econometrics* (2013), 1–38.

[156] Lehalle, C.-A. Market microstructure knowledge needed for controlling an intra-day trading process. In *Handbook on Systemic Risk*, J.-P. Fouque and J. A. Langsam, Eds. Cambridge University Press, 2013, pp. 549–578. Cambridge Books Online.

[157] Lehalle, C.-A., and Dang, N. M. Rigorous post-trade market impact measurement and the price formation process. *Trading 2010*, 1 (2010), 108–114.

[158] Li, C.-B., and Komatsuzaki, T. Aggregated Markov model using time series of single molecule dwell times with minimum excessive information. *Phys. Rev. Lett. 111* (Aug 2013), 058301.

[159] Liao, S. *Beyond perturbation: introduction to the homotopy analysis method*. CRC press, 2003.

[160] Liao, S. On the homotopy analysis method for nonlinear problems. *Applied Mathematics and Computation 147*, 2 (2004), 499–513.

[161] Liao, S. Notes on the homotopy analysis method: some definitions and theorems. *Communications in Nonlinear Science and Numerical Simulation 14*, 4 (2009), 983–997.

[162] Liao, S. *Homotopy analysis method in nonlinear differential equations*. Springer, 2012.

[163] Liesenfeld, R., Nolte, I., and Pohlmeier, W. Modelling financial transaction price movements: a dynamic integer count data model. *Empirical Economics 30*, 4 (2006), 795–825.

[164] Lillo, F., and Farmer, J. D. The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics 8*, 3 (2004).

[165] Lillo, F., Farmer, J. D., and Mantegna, R. N. Econophysics: Master curve for price-impact function. *Nature 421*, 6919 (2003), 129–130.

[166] Lillo, F., Mike, S., and Farmer, J. D. Theory for long memory in supply and demand. *Phys. Rev. E 71* (Jun 2005), 066122.

[167] Lin, J. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on 37*, 1 (1991), 145–151.

[168] Liu, C.-S. The essence of the generalized Taylor theorem as the foundation of the homotopy analysis method. *Communications in Nonlinear Science and Numerical Simulation 16*, 3 (2011), 1254–1262.

[169] Liu, T., Granger, C. W. J., and Heller, W. P. Using the correlation exponent to decide whether an economic series is chaotic. *Journal of Applied Econometrics 7*, S1 (1992), S25–S39.

[170] Loistl, O., Schossmann, B., and Veverka, A. Tick size and spreads: The case of Nasdaq's decimalization. *European Journal of Operational Research 155*, 2 (2004), 317–334.

[171] Lorenz, J., and Almgren, R. Mean–variance optimal adaptive execution. *Applied Mathematical Finance 18*, 5 (2011), 395–422.

[172] Luccioli, S., Livi, R., and Politi, A. Thermal conduction in classical low-dimensional lattices. *Physics Reports 377*, 1 (2003), 1 – 80.

[173] Luccioli, S., Olmi, S., Politi, A., and Torcini, A. Collective dynamics in sparse networks. *Phys. Rev. Lett. 109* (Sep 2012), 138103.

[174] Luccioli, S., and Politi, A. Irregular collective behavior of heterogeneous neural networks. *Phys. Rev. Lett. 105* (Oct 2010), 158104.

[175] Lyapunov, A. M. The general problem of the stability of motion. *International Journal of Control 55*, 3 (1992), 531–534.

[176] MacKinnon, G., and Nemiroff, H. Tick size and the returns to providing liquidity. *International Review of Economics & Finance 13*, 1 (2004), 57–73.

[177] Madhavan, A., Richardson, M., and Roomans, M. Why do security prices change? A transaction-level analysis of NYSE stocks. *Review of Financial Studies 10*, 4 (1997), 1035–1064.

[178] Magnasco, M. O., Piro, O., and Cecchi, G. A. Self-tuned critical anti-hebbian networks. *Phys. Rev. Lett. 102* (Jun 2009), 258102.

[179] Magnus, J. R., and Neudecker, H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons, 2007.

[180] Mandelbrot, B. The variation of some other speculative prices. *Journal of Business* (1967), 393–413.

[181] Mandelbrot, B., and Taylor, H. M. On the distribution of stock price differences. *Operations research 15*, 6 (1967), 1057–1062.

[182] Mandelbrot, B. B. Comments on:" A subordinated stochastic process model with finite variance for speculative prices," by Peter K. Clark. *Econometrica: Journal of the Econometric Society* (1973), 157–159.

[183] Massey, F. J. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association 46*, 253 (1951), 68–78.

[184] Mastromatteo, I., Tóth, B., and Bouchaud, J.-P. Agent-based models for latent liquidity and concave price impact. *Phys. Rev. E 89* (Apr 2014), 042805.

[185] McKenzie, E. *Discrete variate time series*. Handbook of statistics. Elsevier, 2003, pp. 573–606.

[186] Medeiros, M. C., Teräsvirta, T., and Rech, G. Building neural network models for time series: a statistical approach. *Journal of Forecasting 25*, 1 (2006), 49–75.

[187] Mirollo, R., and Strogatz, S. Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics 50*, 6 (1990), 1645–1662.

[188] Molabahrami, A., and Khani, F. The homotopy analysis method to solve the Burgers–Huxley equation. *Nonlinear Analysis: Real World Applications 10*, 2 (2009), 589–600.

[189] Monteforte, M., and Wolf, F. Dynamic flux tubes form reservoirs of stability in neuronal circuits. *Phys. Rev. X 2* (Nov 2012), 041007.

[190] Moro, E., Vicente, J., Moyano, L. G., Gerig, A., Farmer, J. D., Vaglica, G., Lillo, F., and Mantegna, R. N. Market impact and trading profile of hidden orders in stock markets. *Physical Review E 80*, 6 (2009), 066102.

[191] Münnix, M. C., Schäfer, R., and Guhr, T. Impact of the tick-size on financial returns and correlations. *Physica A: Statistical Mechanics and its Applications 389*, 21 (2010), 4828–4843.

[192] Nayfeh, A. H. *Perturbation methods*. John Wiley & Sons, 2008.

[193] Noether, G. E. Note on the Kolmogorov statistic in the discrete case. *Metrika 7*, 1 (1963), 115–116.

[194] Obizhaeva, A., and Wang, J. Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets 16*, 1 (2013), 1–32.

[195] O'Hara, M., Saar, G., and Zhong, Z. Relative tick size and the trading environment. Tech. rep., working paper, Cornell University, 2013.

[196] Olmi, S., Livi, R., Politi, A., and Torcini, A. Collective oscillations in disordered neural networks. *Phys. Rev. E 81* (Apr 2010), 046119.

[197] Onnela, J.-P., Töyli, J., and Kaski, K. Tick size and stock returns. *Physica A: Statistical Mechanics and its Applications 388*, 4 (2009), 441–454.

[198] Osborne, M. Periodic structure in the brownian motion of stock prices. *Operations Research 10*, 3 (1962), 345–379.

[199] Ott, E. *Chaos in dynamical systems*. Cambridge university press, 2002.

[200] Park, S.-H., and Kim, J.-H. Homotopy analysis method for option pricing under stochastic volatility. *Applied Mathematics Letters 24*, 10 (2011), 1740–1744.

[201] Pascual, R., and Veredas, D. What pieces of limit order book information matter in explaining order choice by patient and impatient traders? *Quantitative Finance 9*, 5 (2009), 527–545.

[202] Perold, A. F. The implementation shortfall: Paper versus reality. *Streetwise: the best of the Journal of portfolio management* (1998), 106.

[203] Plasmans, J., Verkooijen, W., and Daniels, H. Estimating structural exchange rate models by artificial neural networks. *Applied Financial Economics 8*, 5 (1998), 541–551.

[204] Plerou, V., Gopikrishnan, P., Gabaix, X., and Stanley, H. E. Quantifying stock-price response to demand fluctuations. *Phys. Rev. E 66* (Aug 2002), 027104.

[205] Plerou, V., Gopikrishnan, P., Nunes, A. L., Meyer, M., and Stanley, H. Scaling of the distribution of price fluctuations of individual companies. *Phys. Rev. E 60* (Dec 1999), 6519–6529.

[206] Plerou, V., Gopikrishnan, P., and Stanley, H. Quantifying fluctuations in market liquidity: Analysis of the bid-ask spread. *Phys. Rev. E 71* (Apr 2005), 046131.

[207] Polyanin, A. D., and Manzhirov, A. V. *Handbook of integral equations*. CRC press, 2012.

[208] Ponzi, A., Lillo, F., and Mantegna, R. N. Market reaction to a bid-ask spread change: A power-law relaxation dynamics. *Phys. Rev. E 80* (Jul 2009), 016112.

[209] Potters, M., and Bouchaud, J.-P. More statistical properties of order books and price impact. *Physica A Statistical Mechanics and its Applications 324* (June 2003), 133–140.

[210] Predoiu, S., Shaikhet, G., and Shreve, S. Optimal execution in a general one-sided limit-order book. *SIAM Journal on Financial Mathematics 2*, 1 (2011), 183–212.

[211] Preis, T. Gpu-computing in econophysics and statistical physics. *The European Physical Journal-Special Topics 194*, 1 (2011), 87–119.

[212] Press, W. H. *Numerical recipes in C 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[213] Prokhorov, A. B. Nonlinear dynamics and chaos theory in economics: A historical perspective. *Quantile 4* (2008), 1–27.

[214] Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (1989), 257–286.

[215] Robert, C. Y., and Rosenbaum, M. A new approach for the dynamics of ultra-high-frequency data: The model with uncertainty zones. *Journal of Financial Econometrics 9*, 2 (2011), 344–366.

[216] Roch, A., and Mete Soner, H. Resilient price impact of trading and the cost of illiquidity. *International Journal of Theoretical and Applied Finance 16*, 06 (2013), 1350037.

[217] Roll, R. A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance 39*, 4 (1984), 1127–1139.

[218] Roulston, M. S. Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena 125*, 3 (1999), 285–294.

[219] Rou, I. A dynamic model of the limit order book. *Review of Financial Studies 22*, 11 (2009), 4601–4641.

[220] Russell, J. R. Econometric modeling of multivariate irregularly-spaced high-frequency data. *Manuscript, GSB, University of Chicago* (1999).

[221] Russell, J. R., and Engle, R. F. A discrete-state continuous-time model of financial transactions prices and times. *Journal of Business & Economic Statistics 23*, 2 (2005), 166–180.

[222] Rydberg, T. H., and Shephard, N. Dynamics of trade-by-trade price movements: decomposition and models. *Journal of Financial Econometrics 1*, 1 (2003), 2–25.

[223] Sánchez, D. A. Calculus of variations for integrals depending on a convolution product. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze* (1964), 233–254.

[224] Sánchez, D. A. On extremals for integrals depending on a convolution product. *Journal of Mathematical Analysis and Applications 11* (1965), 213–216.

[225] Sánchez, D. A. Some existence theorems in the calculus of variations. *Pacific Journal of Mathematics 19*, 2 (1966), 357–363.

[226] Scheinkman, J. A., and LeBaron, B. Nonlinear dynamics and stock returns. *Journal of Business* (1989), 311–337.

[227] Schied, A., and Schöneborn, T. Risk aversion and the dynamics of optimal liquidation strategies in illiquid markets. *Finance and Stochastics 13*, 2 (2009), 181–204.

[228] Schurmann, T., and Grassberger, P. Entropy estimation of symbol sequences. *Chaos: An Interdisciplinary Journal of Nonlinear Science 6*, 3 (1996).

[229] Slanina, F. Mean-field approximation for a limit order driven market model. *Phys. Rev. E 64* (Oct 2001), 056136.

[230] Smith, A., Naik, P. A., and Tsai, C.-L. Markov-switching model selection using Kullback–Leibler divergence. *Journal of Econometrics 134*, 2 (2006), 553–577.

[231] Smith, E., Farmer, J., Gillemot, L., and Krishnamurthy, S. Statistical theory of the continuous double auction. *Quantitative Finance 3*, 6 (2003), 481–514.

[232] Sompolinsky, H., Crisanti, A., and Sommers, H. J. Chaos in random neural networks. *Phys. Rev. Lett. 61* (Jul 1988), 259–262.

[233] Teräsvirta, T., Van Dijk, D., and Medeiros, M. C. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting 21*, 4 (2005), 755–774.

[234] Thakur, V., Azad, R., and Ramaswamy, R. Markov models of genome segmentation. *Phys. Rev. E 75* (Jan 2007), 011915.

[235] Timmermann, A. Moments of Markov switching models. *Journal of Econometrics 96*, 1 (2000), 75–111.

[236] Tong, H., and Lim, K. S. Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society. Series B (Methodological)* (1980), 245–292.

[237] Torre, N. BARRA market impact model. `www.barra.com/research/newsletter/nl165mim.aspx`, 1998.

[238] Tóth, B., Lempériére, Y., Deremble, C., De Lataillade, J., Kockelkoren, J., and Bouchaud, J.-P. Anomalous price impact and the critical nature of liquidity in financial markets. *Phys. Rev. X 1* (Oct 2011), 021006.

[239] Tumminello, M., Lillo, F., and Mantegna, R. N. Kullback-Leibler distance as a measure of the information filtered from multivariate data. *Phys. Rev. E 76* (Sep 2007), 031123.

[240] Turkyilmazoglu, M. A note on the homotopy analysis method. *Applied Mathematics Letters 23*, 10 (2010), 1226–1230.

[241] Urdapilleta, E. Onset of negative interspike interval correlations in adapting neurons. *Phys. Rev. E 84* (Oct 2011), 041904.

[242] Verma, A., Schug, A., Lee, K. H., and Wenzel, W. Basin hopping simulations for all-atom protein folding. *The Journal of chemical physics 124*, 4 (2006), 044515.

[243] Vreeswijk, C. V., and Sompolinsky, H. Chaotic balanced state in a model of cortical circuits. *Neural computation 10*, 6 (1998), 1321–1371.

[244] Wang, C., and Pop, I. Analysis of the flow of a power-law fluid film on an unsteady stretching surface by means of homotopy analysis method. *Journal of non-newtonian fluid mechanics 138*, 2 (2006), 161–172.

[245] Waterfall, J., Casey, F., Gutenkunst, R., Brown, K., Myers, C., Brouwer, P., Elser, V., and Sethna, J. Sloppy-model universality class and the Vandermonde matrix. *Phys. Rev. Lett. 97* (Oct 2006), 150601.

[246] Wazwaz, A.-M. *Linear and nonlinear integral equations: methods and applications*. Springer, 2011.

[247] Weinberger, E. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological cybernetics 63*, 5 (1990), 325–336.

[248] Weinberger, E. Local properties of Kauffman's N - k model: A tunably rugged energy landscape. *Phys. Rev. A 44* (Nov 1991), 6399–6413.

[249] Weron, R. Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A: Statistical Mechanics and its Applications 312*, 1 (2002), 285–299.

[250] Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M., and Vettorazzo, M. Relation between bid-ask spread, impact and volatility in order-driven markets. *Quantitative Finance 8*, 1 (2008), 41–57.

[251] Yabushita, K., Yamashita, M., and Tsuboi, K. An analytic solution of projectile motion with the quadratic resistance law using the homotopy analysis method. *Journal of Physics A: Mathematical and theoretical 40*, 29 (2007), 8403.

[252] Yao, C. and Ye, M. Tick Size Constraints, High-Frequency Trading, and Liquidity. `http://ssrn.com/abstract=2478216`, 2014.

[253] Zaccaria, A., Cristelli, M., Alfi, V., Ciulla, F., and Pietronero, L. Asymmetric statistics of order books: The role of discreteness and evidence for strategic order placement. *Phys. Rev. E 81* (Jun 2010), 066101.

[254] Zarinelli, E., Treccani, M., Farmer, J., and Lillo, F. Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate. *arXiv:1412.2152v1* (Dec. 2014).

[255] Zhu, S.-P. An exact and explicit solution for the valuation of american put options. *Quantitative Finance 6*, 3 (2006), 229–242.

[256] Zucchini, W., and MacDonald, I. L. *Hidden Markov models for time series: an introduction using R*. CRC Press, 2009.