Trading and Price Diffusion:  
Stock Market Modeling Using the Approach of Statistical Physics

Ph.D. thesis

László Gillemot  
Budapest University of Technology and Economics  
Department of Theoretical Physics

Supervisors:  
Dr. Doyne Farmer  
Dr. János Kertész

Budapest

November 2, 2006
Acknowledgment

I would like to thank Doyne Farmer and János Kertész for their supervision and invaluable support. I have been pleased to collaborate with the members of our market research team and I would like to express appreciation (in alphabetical order) to Marcus Daniels, Francois Ghoulmie, Giulia Iori, Supriya Krishnamurthy, Fabrizio Lillo, Makoto Nirei, Paolo Patelli, Eric Smith, Anindya Sen, Szabolcs Mike and Ilija Zovko. I am also grateful to colleagues at the Santa Fe Institute, to Stathis Avdis, Samuel Bowles, Murray Gell-Mann, Constantino Tsallis, John Miller, and Laura Ware for all the useful discussions and support. I thank the Santa Fe Institute for the fellowship and research opportunity. I greatly appreciate the help, support and patience from my parents, sisters, and brothers-in-law without which this study could never have been completed. I thank my friends for encouragement and entertainment, which also contributed to the success of this work.
# Contents

1 Introduction ........................................... 1
  1.1 Motivation ........................................ 1
  1.2 About the economy in general .................... 3
    1.2.1 What is the economy? ......................... 3
    1.2.2 Complex systems and the economy ............ 4
    1.2.3 What can physics do for economics? ........ 5
    1.2.4 The basics of investments ................... 7
  1.3 Components of financial markets ................. 10
    1.3.1 Asset types .................................. 11
    1.3.2 Market structures ............................ 13
    1.3.3 Trader types and their motivation .......... 14
  1.4 Price formation .................................. 15
    1.4.1 Idealized trading ............................. 15
    1.4.2 Supply, demand and market impact .......... 16
    1.4.3 Market efficiency ............................ 17
    1.4.4 Price diffusion ............................... 20
    1.4.5 Stylized facts ................................. 21
    1.4.6 Limit order based trading mechanism ....... 22

2 A simplistic limit order based trading model .... 25
  2.1 Background ...................................... 25
    2.1.1 Motivation ................................... 25
    2.1.2 The model .................................... 27
    2.1.3 Summary of prior work ....................... 29
  2.2 Predictions of the model ......................... 30
    2.2.1 Dimensional analysis ........................ 30
    2.2.2 Varying the granularity parameter $\epsilon$ ... 36
    2.2.3 Varying tick size $dp/p_c$ ................. 45

3 On the origin of large price changes .......... 49
  3.1 Background ...................................... 49
    3.1.1 What causes returns? ....................... 50
    3.1.2 Review of previous work .................... 51
  3.2 Data ............................................. 52
  3.3 Fluctuations in liquidity drive the tails ....... 54
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.1</td>
<td>Large returns are not caused by large orders</td>
<td>54</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Volume dependence of mean market impact</td>
<td>57</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Correlations between order size and liquidity</td>
<td>57</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Liquidity fluctuations drive price fluctuations</td>
<td>60</td>
</tr>
<tr>
<td>3.4</td>
<td>Granularity of supply and demand</td>
<td>60</td>
</tr>
<tr>
<td>3.5</td>
<td>Limit orders and cancellations</td>
<td>63</td>
</tr>
<tr>
<td>3.6</td>
<td>Correlations in occupied sites</td>
<td>67</td>
</tr>
<tr>
<td>4</td>
<td>Stylized facts on longer time scales: where do they come from?</td>
<td>70</td>
</tr>
<tr>
<td>4.1</td>
<td>Background</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>Data</td>
<td>72</td>
</tr>
<tr>
<td>4.3</td>
<td>Volatility under alternative time clocks</td>
<td>73</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Alternative time clocks</td>
<td>73</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Volatility in transaction time</td>
<td>74</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Correlations between transaction time and real time</td>
<td>75</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Correlations between volume time and real time</td>
<td>75</td>
</tr>
<tr>
<td>4.4</td>
<td>Distributional properties of returns</td>
<td>79</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Holding volume or transactions constant</td>
<td>79</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Comparison to work of Ane and Geman</td>
<td>81</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Implications for the theory of Gabaix et al.</td>
<td>84</td>
</tr>
<tr>
<td>4.5</td>
<td>Long-memory</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>Conclusions</td>
<td>92</td>
</tr>
<tr>
<td>A</td>
<td>Theoretical analysis of the simplistic model</td>
<td>96</td>
</tr>
<tr>
<td>A.1</td>
<td>Summary of analytic methods</td>
<td>96</td>
</tr>
<tr>
<td>A.2</td>
<td>Characterizing limit-order books: dual coordinates</td>
<td>97</td>
</tr>
<tr>
<td>A.3</td>
<td>Frames and marginals</td>
<td>99</td>
</tr>
<tr>
<td>A.4</td>
<td>Factorization tests</td>
<td>101</td>
</tr>
<tr>
<td>A.5</td>
<td>Comments on renormalized diffusion</td>
<td>102</td>
</tr>
<tr>
<td>A.6</td>
<td>Master equations and mean-field approximations</td>
<td>103</td>
</tr>
<tr>
<td>A.6.1</td>
<td>A number density master equation</td>
<td>103</td>
</tr>
<tr>
<td>A.6.2</td>
<td>Solution by generating functional</td>
<td>105</td>
</tr>
<tr>
<td>A.6.3</td>
<td>Screening of the market-order rate</td>
<td>107</td>
</tr>
<tr>
<td>A.6.4</td>
<td>Verifying the conservation laws</td>
<td>107</td>
</tr>
<tr>
<td>A.6.5</td>
<td>Self-consistent parametrization</td>
<td>108</td>
</tr>
<tr>
<td>A.6.6</td>
<td>Accounting for correlations</td>
<td>110</td>
</tr>
<tr>
<td>A.6.7</td>
<td>Generalizing the shift-induced source terms</td>
<td>110</td>
</tr>
<tr>
<td>A.7</td>
<td>A mean-field theory of order separation intervals: the Independent</td>
<td>111</td>
</tr>
<tr>
<td>A.7.1</td>
<td>Asymptotes and conservation rules</td>
<td>114</td>
</tr>
<tr>
<td>A.7.2</td>
<td>Direct simulation in interval coordinates</td>
<td>116</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The purpose of this introduction is to explain the motivation behind the economy oriented research and to define the place of economics within science. We also summarize the most important concepts related to the economy and to financial markets, furthermore we confirm the relevance of the statistical physics approach. In order to avoid the common mistake often made in interdisciplinary research, we make an extra effort to introduce the discipline of economics to help the reader to gain a comprehensive overview. In the later chapters this will be extended with more technical background information.

1.1 Motivation

It is a commonly accepted notion that the economy is a self organizing adaptive evolutionary system, as it ought to be. People constantly compare and evaluate goods, services, contracts, and companies by trading them. In other words, everything that is traded is being evaluated simultaneously. Competition, such as that among companies, is an essential part of the economy, and the rise and fall of corporations are a natural part of the evolutionary process that keeps the whole system evolving [1]. This view is confirmed by history as well: the suppression of competition and trading restrictions in the centrally planned economic model essentially led to the fall of these systems. In a market economy, in contrast, even trading of shares of companies is essential as it facilitates the overall economic equilibrium.

The market economy has already proven its vitality, however, it is commonly believed that stock markets tend to be oversensitive and vulnerable to an extent that no simple theory can explain. Values of stocks fluctuate on a much faster time scale than can the values of their respective companies. Furthermore, it is known that stock market crashes may happen with no good reason, i.e., even when there is no bad news arrives in the market. This is clearly harmful for the economy because it implies excessive risk, and that is costly. Such events are now suspected of being driven by the structure of trading and institutional rules rather than by factors that are directly related to the companies themselves. However, the origin of these disadvantageous characteristics in price formation is not yet
known. Understanding this problem could pay off by providing the knowledge to improve the institutional rules in order to significantly reduce vulnerability and implied risk in the economy. That is in the interests of both the people and the institutions.

On the other hand, economics has many interesting scientific points of view. Perhaps the strongest motivation for physicists to become interested in financial economics is a set of so-called stylized facts. This concept includes three major non-trivial statistical properties of the price time series of financial assets that surprisingly can be observed for different assets of different asset types, across different markets, for different time periods, and for different time scales. But how could we explain the robustness of the stylized facts if the economy was exclusively driven by news directly related to the economy and by erratic acts of people, while the economy is evolving and institutional rules are changing all the time? The explanation seems to be beyond the scope of the conventional neoclassical approach.

Statistical physics gives examples for similar phenomena. Characteristics of macroscopic quantities of a many-particle system are determined by the particular properties of the particles and by the interactions among them. However, the macroscopic behavior is usually not sensitive to many microscopic details, only to a few dominant ones. The robust presence of the stylized facts in the financial data foreshadows the existence of a profound microstructure in the economy, which accounts for these particular macroscopic phenomena.

For a physicist, it is also very suggestive that in economic data there is a large number of quantities which show power-law characteristics. Although many processes exist which can generate power-laws, they might be rooted in critical behavior in the statistical physics sense.

Furthermore, the breakthroughs in computer technology over the last couple of decades have made it possible to model and simulate complex systems. Physicists’ experience in simulating complex physical systems can prove useful in economic research as well. At the same time, improvements in information technology have not only had an invaluable impact on the technology of trading in financial markets, they have boosted data gathering and enhanced possibilities in data analysis. The number of transactions per unit time is still growing rapidly as trading becomes more automated, and larger and larger fractions of the recorded data are well detailed. Numerical methods and simulations are now essential tools for solving the non-trivial, non-linear models. There are still numerous open questions in economics, but we now have the data and the tools to answer them.

Finally, many of the conventional economics schools tend to be excessively conservative. They often prefer to make extreme, unrealistic assumptions such as perfectly rational agents, perfectly efficient markets, or Gaussian distributed returns, in order to keep problems more tractable analytically. However, these assumptions are usually far from reality and bear little practical significance. Not surprisingly, industrial experts use only a fraction of the academic results, and this leaves an empty space between practice and theory. Physicists can contribute toward filling this gap, and adding value by considering new aspects,
taking advantage of quantitative modeling skills, and using methods learned in 
physics.

1.2 About the economy in general

1.2.1 What is the economy?

The economy is an evolutionary social system which is devoted to serving the people’s material needs and improving their living environment and conditions. It is inevitably present in everyday life. We take part in elementary economic processes when we purchase goods, services, and when we earn our salaries. Just as in any evolutionary system, efficiency is in the spotlight of concern. Higher efficiency forms the basis of better welfare, thus increasing efficiency is in the interest of the people in general. However, the economy is a complex system with many non-trivial phenomena, thus it is usually not obvious what the results of the various impacts would be at the global level. In order to be able to improve efficiency though feasible institutional regulations, we first need to understand better how the economy works.

Higher efficiency is typically achieved by better exploitation of work resources. The organization of production processes and specialization of work resources are the major ingredients of efficiency. Without these, everyone would be restricted to producing everything themself. They would obviously be unable to produce the vast majority of goods and services (e.g., groceries, phone calls, cars, etc.) we depend on in everyday life. However, specialization is only possible if we let people swap their goods and services, that is, if we let them trade. Hence, trading is perhaps the most essential mechanism in the economy.

Indeed, we may say that the history of the economy began when two people swapped goods for the first time. The complex structure of the economy has evolved since then. The mechanism of free trading is robust because it is in the interests of both the seller and the buyer, or at least they believe so. Otherwise, the party which finds the deal disadvantageous simply rejects it and no transaction takes place. Note that the value is subjective, as it must be different for the two parties. However, the transaction is completed at a given price, therefore, price is objective, even if it may change from transaction to transaction.

Trades can be classified according to the underlying asset. As a buyer, an individual may take two essentially different actions. On one hand she may buy some consumer products (goods, services). This spending reflects the utility (i.e., personal preferences) of the individual as she picks those items that she evaluates most for the available resources. At the scale of the economy, all the preferences of the people accumulate and thus reflect the aggregated demand of consumption items to which the production adapts. The primary goal of the economy is to serve this demand as well as possible, and hence improve the welfare of the individuals\(^1\).

On the other hand, one can make an investment, that is, tie up money for expected future payoffs, which are greater in total than the investment. This is

\(^1\)The term welfare also implies basic needs such as food.
possible because the money is invested (directly or indirectly) into facilities that serve consumer goods production and make a profit. A fraction of the profit is paid back to the investor, proportional to the amount of invested capital. Then, this income can be spent on consumer goods or reinvested, or perhaps combined. However, all the investments only make sense if they eventually get spent (directly or indirectly) on consumption, increasing welfare\(^2\).

Although the ultimate goal of the economy is to serve consumption, investments are just as important because they are part of the process of efficient production. Capital may be used to establish a new company which can contribute (directly or indirectly) to consumer product production. Financial markets add value by making it possible for the investors to sell (a fraction of) their company for a good price (market price), thereby freeing up some of their capital. This capital then may be used to start another new firm. As this cycle repeats, the economy continues growing.

Furthermore, financial markets provide an efficacious environment for evaluating and comparing investments. A financial investment (i.e., holding financial assets) can be made to realize future earnings. An example that typically involves all the people living in a modern economy is pension funds. Although pension service may be managed by the government –typically very inefficiently– for historical reasons, the most effective way to manage pension is through investments made in financial markets.

Note that there are numerous possible reasons why an individual would decide to make an investment rather than to spend it on consumption. Indeed, there is a competition between these two possibilities: one can decide either to buy some consumer product instantly or to ensure having more money to spend in the future.

### 1.2.2 Complex systems and the economy

Systems built of strongly interacting constituents are often appear to be complex. The behavior of a complex systems is much richer as some phenomena emerge which could not be expected simply by considering the independent set of constituents. There are many features and mechanism that can create or increase complexity. Let us review the most common ones next.

It is a recent trend that interdisciplinary areas increasingly attract the attention of scientists. This is a straightforward consequence of the fact that problems occurring in the real world are rarely dominated by a single factor and these factors are often from areas which are classified into separate scientific disciplines. Many systems are built of subsystems which are nested into each other. For example, in order to understand the speed-wise performance of the internet, one has to consider the network characteristics of the internet and not only the computer technology and computer science aspects [2].

Complexity can be observed in many pure systems as well. The non-uniform coupling in a many particle system can also establish complexity. Physicists first
think of a spin glass as an example perhaps, but this is also the case of a living body made of cells or a processor made of transistors.

On the other hand, even simple but non-linear dynamical systems can exhibit complex behavior patterns at the edge of chaos, as, for example, Hofstadter’s butterfly phenomenon demonstrates [3].

Evolution is a robust mechanism that can lead to emergent phenomena as well and, therefore, drastically increase complexity.

The economy as a whole is a complex system. At first glance it may seem to be surprising that the economy is driven by humans’ decisions, yet we have a quite limited understanding of it. Nevertheless, we see how complexity emerges as we can identify the above mentioned structures and mechanisms. Let us next consider a short list of examples.

In the economy, we find different nested subsystems rooted in the psychology of humans, game theory, networks of business relations, technology, and other conventional economic fields. On the other hand, most of the processes in the economy are non-linear. For instance, decision making is an essential mechanism in the economy, and it is a repeated non-linear process. This alone is sufficient for the emergence of chaos. Furthermore, the economy is a system with a large number of interacting elements as the individuals and the corporations trade with each other. The consequences of the latter are non-trivial and there has been little emphasis on this fact. Finally, it is obvious that we can observe evolution in the economy at different levels. There is selection: companies which cannot sell their products to the customers go bankrupt, investors who lose their inventories are not able to invest any more, etc., so they fall out of the play. The economy manifests adaptation as companies update their profiles and improve their structures according to economic conditions, change their employees, and so on.

Nature, as well as artificial and abstract systems, provides an uncountable number of examples of complex systems. However, it is impossible to fully track the behavior of such systems analytically. The analysis typically requires computation power which is just becoming feasible recently. The rapidly developing computer simulations, as well as the ability to process large amounts of data are boosting scientific development in economics similarly to other complex science fields.

1.2.3 What can physics do for economics?

There are several distinct reasons why physicists consider finance and economics. First of all, physics provides a good training in analyzing, understanding and modeling dynamical processes and complex systems. Using examples from nature, physicists learned straightforward problem solving by separating dominant and negligible factors. Many mathematical and physical tools were developed over the centuries to support this task. Physics assigns equal importance to theory and to experiments; and there is no way to isolate them from each other.

Some of the methods from physics are directly applicable to financial problems. A good example is the random matrix theory, which proved its use in
portfolio optimization theory [5, 6].

Demand is another important reason. Most of the conventional academic theories tend to be quite conservative, and they have little relationship to the modern-day industry. Over the past decades the gap between academia and industry has increased even more, mainly because of the rapid development in industry, where new challenges continuously occur. Nowadays, most of the modern quantitative studies are born in industrial research centers rather than in academic ones. However, these works are usually limited to profit-oriented problems and many globally and scientifically interesting questions remain open.

There are even more concrete and profound reasons why physics can contribute to economics. Within the scope of physics, it is statistical physics that is most likely to be related to other scientific fields. The main reason is that it is a discipline that is not exclusively applied to physical systems. It describes systems which are composed of a large number of interacting units. Beyond its physical content, statistical physics provides a framework for dealing with such systems in general. Using physical examples we learned the concept of microscopic and macroscopic levels of description, fluctuations, intensive and extensive quantities, entropy, temperature, phase transitions, critical behavior, universality, direction of processes, reversibility and irreversibility, collective behavior, etc., which are specific to statistical physics, that is, to large systems of interacting particles. In this framework many new tools and concepts were developed and proven useful, such as theories of phase transition, scaling, and renormalization group theory. Statistical physics has already led to significant contributions in the understanding of traffic systems [7, 8, 9], granular materials [10], meteorology [11, 12], biology [13, 14], and many other scientific fields.

Social sciences deal with large numbers of people by definition, but there has been little focus on the high cardinality. Indeed, economics is yet another area in which the approach of statistical physics turns out to be fruitful. In support of this statement let us consider a few factors that play lead roles in financial markets\(^3\). (1) Economy related information, i.e., news arrives in the market. This is the most direct and probably the most understood effect. For example, a financial report of a company, a change in the government’s monetary policy (e.g., taxing), an introduction of a new technology or a natural disaster are factors which greatly affect the economy. These are fundamental effects that constitute the basis of the neoclassical economics. (2) Behavioral economics claims that psychology plays an important role because decisions in business are made by human beings. Modeling the human factor is very challenging. In the literature one finds extreme approaches, almost exclusively. Conservative branches of the research consider traders perfectly rational, while some others say they are well modeled by zero intelligence. The truth perhaps lies somewhere in the middle, and rationality could be characterized by a wide distribution. (3) Traders want profit. No doubt, is the primary motivation. However, it is not always clear how to achieve profit, and even more unclear what is the cost (e.g., risk) of the profit.

Many of the conventional economic theories would finish the list here and

\(^3\)This could be generalized for the whole economy, but we restricted our consideration to financial markets for the sake of simplicity.
would try to deduce everything from these three factors. However, considerations motivated by other disciplines can contribute new points of views. Physicists find the following features interesting: (4) Large numbers of people and companies trade among each other, (5) completing many transactions of (6) many assets. Furthermore, (7) the characteristics of the network of traders and trades are expected to be relevant to an understanding of the economy.

From a dynamical aspect, this is suggestive of how a trading-driven economy can be analogous to an interacting many-particle physical system. In the analogy, point 4 corresponds to the particles, whereas points 2-3, 5-7 are analogous to the properties and related quantities of the elementary interaction (or collisions) between the particles. A fraction of the news in point 1, which is about the global economy corresponds to external forces, the rest may also be considered as a part of the interaction.

The scope of this study is limited to the analysis of the mechanism of trading, including its general properties and its consequences. In the example above, this corresponds to the analysis of the elementary interaction.

On the other hand, some knowledge from statistical physics can also be applied to certain aspects of the phenomenology of economics. Trade prices of the financial assets exhibit diffusive behavior for a profound reason. Additionally, a great amount of data analysis provided evidence for anomalies in price diffusion, the cause of which is not yet known. Experience with real physical systems may contribute to the understanding of this economic problem. We will provide a detailed introduction for price formation in Section 1.4.

1.2.4 The basics of investments

In our study we analyze price formation based on financial market data, therefore, it is essential to review the most basic concepts related to investments. In order to understand how investment opportunities are evaluated and compared when investment decisions are to be made, we need to introduce the notion of expected return and risk.

We know intuitively that $1 today is more valuable than $1 tomorrow. If one has $1 today, she might simply hold it for a day and she will end up with the same $1 tomorrow with no effort. However, she has the option to do something with that $1 today, such as to take it to the bank and earn the interest by the next day. When we make an investment we expect a series of future payoffs. Obviously, we expect more money to be returned in total as we also want the time value of the invested money to be paid.

There are two quite different ways to make an investment. One can finance a project, that is, spend money to fund (develop) a firm in order to make (increase) profit. The investors of the project are entitled to the profit. The cost of the investment is the cost of the development. Alternatively, one can buy financial assets at financial markets which are expected to pay off in the future in some way. The price of a financial investment is the market price of the corresponding securities, that is the price at which the asset can be bought at financial exchanges.

---

4Securities are the ownership certification documents of the financial assets, thus the word...
However, both investment types can be characterized by their cashflows, that is, by the series of future payoffs.

Cashflows may have very different actual payoff amounts, times, and frequencies without one being clearly better than the other. To compare investments it is useful to convert cashflows into returns. When we make an investment of an amount of money $S_0$ at time $t_0$ that generates a cashflow $s_i$ with payment times $t_i$, we first solve the equation

$$S_0 = \sum_{\{i\}} s_i e^{-\alpha(t_i - t_0)}$$  \hspace{1cm} (1.1)

for $\alpha$. Although $\alpha$ is the most obvious first order characterization of the performance of an investment, the financial practice tends to use the equivalent return instead. It is written as $r_T = e^{\alpha T} - 1$, where $T$ time scale can be arbitrary, but the most often used value is one year. The value of return is expressive: provided an investment generates return $r_T$, the profit on the invested $S_0$ money over time $T$ will be $\Delta S_T = S_0 r_T$.$^5$ If one invests the same amount of money $S_0$ for the same time span, the higher the return is, the more profit is earned. Note that $r$ must be positive for a profitable investment.

As the cashflow cannot be obtained only from the return uniquely, there are higher order characteristics of a cashflow, such as the second order duration and the third order convexity. However, these quantities are not of concern in this study.

Not only project investments, but financial investments are strongly linked to production as well. When we buy a share of a stock (e.g., a share of IBM), we acquire a partial ownership of the company. If the total number of shares issued is $N$, then one share represents $1/N$ fractional ownership and entitles us to $1/N$ fraction of the profit of the company in the future, as long as we hold the share. This may be realized in two different ways. A part of the profit may get directly reinvested in the company to improve production facilities, which increase the future expected profit, resulting in an increase of the price of the stock. The rest of the profit may be paid to the shareholders as dividends. Generally speaking, the profit on a stock investment is the total of the stock price increment and the dividends.$^6$ The profit is realized when the shareholder sells the shares, that is, when the investment is entirely converted into a cashflow.

Note that the return depends on the initially invested money and on subsequent payoffs, as reflected in Eq. 1.1. Thus, at the time when the investment is made, the investor may not know the actual amounts of payoffs. Indeed, the return is calculated based on cashflow expectations, and, therefore, the calculation

$^5$"security" is often used in the context of a financial asset.

$^6$Since $r_{1 \text{ year}} \ll 1$ typically, the linearized form $r_T = r_{1 \text{ year}} \frac{T}{1 \text{ year}}$ is often used in cases when $T < 1 \text{ year}$.

$^6$Depending how fast the company grows, two rough classes are differentiated. A growth stock pays a little (or no) dividend and the majority (or all) of the profit is incorporated in the price increment of the stock. There are numerous examples among the fast growing information technology firms. A value stock typically grows very slowly, thus its price changes slowly, however, it tends to pay a high dividend. For example, utility companies tend to be value stocks.
yields an expected return, rather than the actual value which can be calculated only when the investment is completed. However, the profit, and consequently the return, may be estimated based on fundamental analysis, that is, valuation of the firm. This involves estimates about the success of the products, the total costs of production, revenues, technology developments, taxes, government regulations, etc. Obviously, there are always some uncertainties in such calculations. First, a development may result in a different impact than originally expected. That may be blamed on the management of the firm. Second, there is a particular dynamic that maintains uncertainty. The production of a firm depends on input products and services, prices of which may change. Changes in the input costs affect the profit and often the product prices as well. In the chain of production, the input of one company is the output of another one, thus all the companies have uncertainty in their profits. This reminds us of a many-particle system, where fluctuations exist even in equilibrium and belong to the intrinsic characteristics of the system. Thus, we suspect that uncertainty in the profit and cashflow of an investment is always present.

However, due to the uncertainty of the cashflow, the return is characterized by a distribution rather than by one single value. Obviously, the wider the distribution, the more risk is implied in the investment. In order to characterize the magnitude of risk, a large number of different risk measures are developed and being used [18]. We all know intuitively that risk is costly. That is why we prefer to have our own high value properties (such as a car or a house) insured. Paying a predefined small amount of money on a regular basis may be preferred over risking a possible big loss, even if the total cost of the insurance policy is more than the expected loss. In the economy, risk reduces production efficiency and thus translates into cost. The financial investors take this risk into account and they price the investments accordingly. They estimate the risk related to the investment. The higher the risk, the higher expected return they want, that is, they are willing to pay less initial money $S_0$ for the financial investment in the market (see Eq. 1.1).

A financial investment includes possessing and trading financial assets. The set of securities held at a given time is also called a portfolio. Let us assume that all market participants have their preferences for the expected return as a function of risk. Now let us consider the market population averages of these curves, which should be an increasing function as illustrated in Figure 1.1. Since securities are bought in the financial market at market price, all the investors have an equal chance of opening a portfolio position on the average curve. On the other hand, it should not be possible to create a position above the market benchmark, because that would clearly be a better one and all the other investors would try to create such a position. As a result, the average curve would move so that it would incorporate that position. This oversimplified reasoning suggests that all positions should be on the same, average market expected return-risk curve. However, the risk may be defined in several different

\footnote{A portfolio is a set of shares of stocks and may include other financial instruments. For example, 5000 shares of Google and 30000 shares of Apple together is considered a two-stock portfolio.}
Figure 1.1: An illustration of the expected return as a function of risk. The solid curve represents the market average. Those positions which fall under the curve take too high a risk for the given expected return and thus do not meet the benchmark. In contrast, those above the curve are better hedged for the given expected return than the average.

ways and measuring it is often a complex quantitative task. Market participants make different estimates of expected returns also, therefore, the average market return-risk curve is ill-defined. Furthermore, some of the participants are not eager to optimize their risks and they may take positions with excess risk. Nevertheless, this concept is not irrelevant and most of the market participants tend to take a position close to an average curve. Risk management takes effort and it is of main concern to financial institutions. Minimizing risk for a given expected return level is also known as hedging. This may be done using insurance-like contracts, such as options (see later in Section 1.3.1), or by the most commonly used risk management method of diversification. If we create a portfolio that includes more than one stock, the expected aggregated return will be the volume-weighted average of the individual stock returns, however, the aggregated risk will be below the weighted average of the individual stock risks because of the uncorrelated (but at least not non-perfectly correlated) nature of the stock price changes.

1.3 Components of financial markets

In this section we intend to provide a better insight into real financial markets by describing asset types, market structures, and trader stereotypes.
1.3.1 Asset types

Many classes of financial assets are traded in real financial markets [18]. Next, we review the most common ones.

As we mentioned before, companies with more than one owner are divided into shares, which represent fractional ownership of the firm. A company is either private or public, depending on whether its shares are traded over the counter or in an organized stock market. A company may be made available to the public through an initial public offering, after which the stock can be freely traded in financial exchanges. An often-used collective name for the stocks is equity, and the expression equity market is used as a synonym for stock market. The majority of the overall invested money is in equity, thus it is the most fundamental investment asset. Long-term equity investments realize profit from the dividends and from the price increment of the stock.

As we mentioned in the previous section, portfolios of stocks are especially interesting from the point of view of risk management. There are special, market-wide “virtual” portfolios known as stock market indexes. For example, the S&P 500 for the New York Stock Exchange is perhaps the most well known one. The value of an index is calculated from the most recent trade prices of its stocks and they are considered the primary indicators of the state of the stock exchanges. Since indexes of large markets may contain hundreds of stocks (i.e., hundreds of companies), stock portfolios covering the indexes were introduced to the market and they are traded independently as financial products. They are known as exchange traded funds (ETF). For instance, the ETF for S&P 500 is called Spider. ETFs make it easier to manage large portfolios and they optimize transaction costs. The market dynamics guarantee that the price of an ETF and the prices of the composing stocks are highly consistent (see later in Section 1.4). The weights of the stocks in an index are usually proportional to their capitalization, that is, to the total value of the corresponding company, as suggested by the optimal portfolio theory called capital asset pricing model [19].

The most liquid asset in an economy is the money. It may be considered as the ETF for the index of the entire economy. The economy-wide index includes not only financial instruments but anything that is traded in the economy. Because of this high diversity, money is a low risk asset. However, it is not a good investment instrument: it includes all the low and negative returning assets as well. An inventory of money usually is reinvested into other assets, which have positive expected returns. This is how banks can pay an interest on money deposits. On the other hand, one may trade in foreign currency exchanges, where money of different economies is exchanged. The foreign exchanges are very liquid markets which support the import and export of the different countries. The currency speculators add value by providing liquidity while they take risk as well.

A bond is a form of loan contracted by a corporation or a government through financial markets. Unlike bank loans, bonds do not tie the contract to the lender. First, the issuer sells the bonds to the investors in the primary market, typically in large quantities, then the bonds can be freely traded in the secondary market. Bonds pay fixed amounts of money on predefined future dates, that is, a steady
cashflow is contracted at the time when the investment is made. The risk of a bond implies the possibility that the issuer makes the payments with delay, or does not pay the contracted amount. Therefore, the reputation and the financial condition of the issuer is very important. Bonds issued by reputable governments are considered risk-free investments. Even though some bonds may be risk-free from the point of view of the cashflow, they do have risk in their trade price. If we need to sell a bond before its maturity, that is, before the contracted cashflow completes, we may get only a low price for it if bonds with a higher return were issued meantime. The so-called risk-free bonds guarantee their nominal returns only if they are held until maturity. Bonds are the other major investment instruments beside stocks.

Commodities are goods and raw materials which are available in standard quality and form which are traded in organized markets. These are typically energy sources (oil, gasoline, electricity, natural gas), metals (gold, silver, platinum, copper, aluminum, lead, zinc), grains (soy, wheat, corn, oat, rapeseed), live-stocks (cattle, hogs), food ingredients (orange juice, sugar, cocoa, coffee, milk), and other kinds of raw materials (rubber, cotton, lumber, ammonia). Commodity markets bring together the consumers and the producers, ensuring liquidity, and they also provide an environment for the participants to hedge the risk involved in their business.

Derivatives, also known as contingent claims, are contracts in which values depend on the future value of one or more other assets, the so called underlying assets. Originally, commodity markets motivated the introduction of derivatives to source out some of the risk of the producers to the speculators. There is an unlimited number of ways to design a derivative, thus we list only the most common ones below.

A forward contract obligates the contractor to exchange the underlying asset at a given price at a given time in the future. Historically, forwards were invented to hedge the risk of agricultural commodities (e.g., wheat) by enabling the producers to sell their crops at a given price before the harvest. A future is an improved version of the forward contract, which is better adapted to the practice of financial trading.

A call (put) option is the right, but not an obligation, to buy (sell) the underlying asset on a certain date (called the exercise date) at a fixed price called the strike price, or exercise price. On the exercise date, if the market price of the underlying asset is above the exercise price, the owner of the call option can exercise the contract and buy the underlying asset at the strike price. Then the asset may be sold in the market right away to earn the difference. If the price of the underlying asset is below the strike price, the contract is worthless. A call option is, therefore, an insurance policy which protects its owner against a potential increase of the price of a given asset which she intends to buy in the future. Symmetrically, a put option insures a minimal price to its owner for her asset which is to be sold in the future.

Note that all kinds of financial products can be designed and then introduced to financial markets, so many variants and combinations of the above listed asset types exist. However, the only products that remain in the market for a long time
are those that facilitate the operations of a large group of market participants, thus they are traded frequently. This is another example of evolution in the economy.

1.3.2 Market structures

More than a hundred organized financial exchanges exist throughout the world. The most well known stock markets are the New York Stock Exchange, Nasdaq, Chicago Stock Exchange, London Stock Exchange, Frankfurt Stock Exchange, Tokyo Stock Exchange and Stock Exchange of Hong Kong, however, there are many other large markets which are specialized in other kinds of financial products, such as derivatives or commodities. All these markets may seem to have significantly different operational rules, trading methods, and trading technologies, however, all of them are devoted to facilitating trading. In order to demonstrate differences between markets, next we compare the trading structures of the New York Stock Exchange (NYSE) and the London Stock Exchange (LSE). We chose these two markets because we use date form them in later chapters of this study.

Both the NYSE and the LSE have dual market structures consisting of a centralized market and a decentralized (bilateral) exchange. The centralized market is called the downstairs market in New York and the on-book market in London, and the decentralized exchange is called the upstairs market in New York and the off-book market in London. While the corresponding components of each market are generally similar between New York and London, there are several important differences. The downstairs market of the NYSE operates through a specialist system. The specialist is given monopoly privileges for a given stock in return for committing to regulatory responsibilities. The specialist keeps the limit order book, which contains limit orders with quotes to buy or sell at specific prices. As orders arrive, they are aggregated, and every couple of minutes orders are matched and the market is cleared. Trading orders are matched based on order of arrival and price priority. During the time of our study, market participants were allowed to see the limit order book only if they were physically present in the market and only with the permission of the specialist. The specialist is allowed to trade for his own account, but also has regulatory duties to maintain an orderly market by making quotes to ensure that prices do not make large jumps. Although a given specialist may handle more than one stock, there are many specialists, so that two stocks chosen at random are likely to have different specialists.

The upstairs market of the NYSE, in contrast, is a bilateral exchange. Participants gather informally or interact via the telephone. Trades are arranged privately and are made public only after they have already taken place [15].

The London Stock Exchange also consists of two markets (London Stock Exchange, [16]). The on-book market (SETS) is similar to the downstairs market and the off-book market (SEAQ) is similar to the upstairs market. In 1999 57% of transactions of LSE stocks occurred in the on-book exchange and in 2002 this number rose to 62%. One important difference between the two markets is that the on-book exchange quotations are public and are published without any delay.
In contrast, since transactions in the off-book market are arranged privately, there are no published quotes. Transaction prices in the off-book market are published only after they are completed.

The on-book market is a fully automated electronic exchange. Market participants have the ability to view the entire limit order book at any instant (even remotely), and to place trading orders and have them entered into the book, executed, or canceled almost instantaneously. Though prices and quotes are completely transparent, the identities of parties placing orders are kept confidential, even to the two counter parties in a transaction. The trading day begins with an opening auction in order to build initial liquidity. There is a period of 10 minutes leading up to the opening auction when orders may be placed or cancelled without transactions taking place, and without any information about volumes or prices. The market is then cleared, i.e., all the possible transactions are completed, and for the remainder of the day (except for rare exceptions) there is a continuous auction, in which orders and cancellations are entered asynchronously and result in immediate action.

The off-book market (SEAQ) is similar to the upstairs market in the NYSE. The main difference is that there is no physical gathering place, so transactions are arranged entirely via telephone. There are also several other minor differences, e.g., regarding the types of allowed transactions and the times when trades need to be reported.

1.3.3 Trader types and their motivation

Financial markets allow two complementary populations to meet: entrepreneurs, who run some industrial projects in need of funding, and investors (institutions and individuals) who have some money to invest in order to gain a share of the future profit and who are willing to take the related risk. The possibility of trading securities freely enables investors to maintain their investments, by which they express their return and risk related preferences. However, the strategy that investors follow when they allocate their money into securities may be very different. Next we discuss the main trader types present in financial markets.

*Fundamental analysts* make decisions based on corporate evaluations and global economy conditions. They calculate the expected future cashflows of the companies and estimate the related risks. Such tasks involve revenue and cost estimation, taking into account the performance of the entire economic sector, expectation about technology and regulation changes, etc., in order to determine the *fundamental price* of the stock. A fundamental analyst may find that the current market price of a given stock is under or over estimates of the fundamental price, then she can long or short his position, respectively. Investments based on fundamental analysis typically follow long-term (several month) strategies.

A *hedger* tries to minimize the risk implied in the financial investment for a given level of expected return, using various risk management methods and financial instruments (e.g., options).

A *market maker* has the very important role of providing liquidity to the market by offering to buy and sell large quantities of a security. The offered buy
price is lower than the offered sell price, thus market makers earn the bid-ask spread defined as the price difference between the sell and buy orders. In some markets the market maker is an institutionally designated role with rights and obligations (for example, the specialist in the New York Stock Exchange), in some other markets it is a role that any trader can take on voluntarily, but not exclusively.

Quantitative analysts (or quants) develop trading strategies based on statistical patterns of prices in historical data. They typically ignore economy related news (or use it only in a statistical sense). They may use excessive amount of trading mechanism related data. Also, they may take advantage of the inconsistencies in the pricing of related financial products (such as derivatives and the underlying assets), or between different markets. Quantitative analysts typically trade based on short-term strategies (shorter than a couple of days).

Any market participant may be considered a speculator who strategically takes a short-term position to take advantage of a non-commonly or statistically expected change in asset prices. Speculation can be based on fundamental considerations (for example, based on a macroeconomic change due to an election) or on historic data analysis. The common judgment of speculators tends to be quite bad, however, they do add value by taking risk and providing liquidity to the market.

Note that traders may play more than one role.

1.4 Price formation

1.4.1 Idealized trading

In the later parts of this study we analyze stock trading data and aim to model stock markets. The main reason we restrict our discussion to financial markets is that only they have detailed and complete data sets of trading records over many years, which makes it possible to create good statistics. But there are other reasons as well. Modern electronic stock exchanges are almost perfect realizations of ideal markets. The shares of a stock are truly identical, there are no quality differences, which cannot be said about most of the consumer products. Due to centralization, the information of supply and demand (but at least the best quotes) are available in real-time to the public\(^8\), so there is no problem with partial trading information. Furthermore, trading is done continuously during the opening hours, unlike certain commodity markets, which obey more restricted institutional rules. They also tend to be very liquid, which is expected to facilitate relaxation to equilibrium. The trading rules are well defined, and possible transactions are completed practically immediately. These are all important ingredients of ideal trading conditions, and thus the stock markets manifest the most idealized trading in the real world.

\(^8\)Data access may imply some fees and other additional costs.
1.4.2 Supply, demand and market impact

In Section 1.2.4 we have featured the factors that affect the asset prices, how stock prices are estimated. Nevertheless, the price of an asset is what it sells for. Indeed, the transactions set the prices, which are up to the seller and the buyer. The trading mechanism does not incorporate the fundamental price directly. The price that the seller and the buyer agree on greatly depends on the supply and the demand, which include all the other competing bids and offers in the market.

In the economy, in general, there are various possible reasons that motivate trading regarding both consumption and investment. However, there is a robust dynamo of trading which is independent from the motivation. Roughly speaking, the supply of a given asset (or product) is the quantity that would be sold at a given price if there were enough buyers and vice versa, the demand is the quantity that would be bought at a given price if there were enough sellers at that price. Supply and demand are obviously functions of the price: a seller who is willing to sell an asset at a given price would obviously sell it more expensively too, that is, the supply increases with the price. In contrast, a buyer who is willing to buy an asset at a given price would obviously buy it for less as well, thus the demand decreases with the price. In equilibrium, supply and demand are equal, the corresponding price is the equilibrium price.

Suppose that trades complete at a non-equilibrium price where supply and demand are not in balance, by definition. In such cases the following simple dynamics take place. If the price exceeds the equilibrium price, the demand will not meet supply and sellers need to drop their prices to be able to sell the desired quantity. If the price is below the equilibrium price, the demand will exceed the supply, so sellers can raise their price (which is in their interest) and still be able to sell the quantity they want. In both cases the price moves toward the equilibrium price, at which the supply and demand become balanced.

Consequently, we affect the price when we take part in the trading. When we intend to sell a given asset we increase the supply, thus the price will tend to go down. For buying it works vice-versa, that is, we increase the demand and the price will tend to go up. This effect is called market impact or price impact. All the traders have impacts on the price, and the aggregated supply and demand define the final transaction price. Note that market impacts always affect the price so it moves in the direction which is less preferred by the action taker. If one finds the current market price high (e.g., a fundamentalist believes it is above the fundamental value), then it is profitable to sell, but selling tends to push the price down making the selling a somewhat less good deal. And vice-versa, buyers always want to get a lower price, however, their presence in the market raises the price. Consequently, the market price should move toward the equilibrium price due to the negative feedback. Note that there is trading even in equilibrium, similarly to a physical gas where there are also collisions in thermal equilibrium.

Although the above sketched conventional picture captures the basics of the dynamics, it is oversimplified. The price formation is driven by the unveiled supply and demand, which is merely a small fraction of the total supply and demand. Furthermore, the unveiled supply and demand exhibit large and fre-
quent fluctuations, as does the transaction price, consequently. There is a range of prices (rather than a just a single value) that could satisfy the assumptions of the equilibrium price according to the definition above. In order to provide a realistic, precisely defined price formation mechanism, we will introduce the formalism of limit order book based trading in Section 1.4.6. It is also the appropriate framework for defining relevant quantities, such as supply and demand, liquidity, market impact, etc., as well as for revealing relationships between them.

1.4.3 Market efficiency

In Section 1.2.4 we briefly discussed how stocks are evaluated according to fundamental analysis. In Section 1.4.2 we roughly summarized the basics of trading-driven price dynamics. Now we aim to explain how these are linked, that is, why trade prices should be close to fundamentally expected prices.

Note that if the trade price of an investment (e.g., a stock) is always the same as the fundamental price, then the only way to make a profit would be to buy and hold the securities for the entire time of the investment. The price would change as the fundamental price changes, which is subject to news arrival causing changes in the expected return and risk. However, the expected return would be positive for a long-term investment if the corresponding company makes a profit from its production. In contrast, if one expects predictable price movements due to particular reasons (e.g., private information) then there is an excess profit opportunity, which can be exploited by intermediate trading.

As described in Section 1.4.2, the market impact of a trade always moves the price so it is less preferred by the trader. Indeed, this property of price dynamics expresses that the exploitable excess profit resources are limited. Even if a trader has private information about an expected future price movement, she still cannot make unlimited profit. As traders exploit the resources they weaken them because of their market impact. Since traders are very much interested in exploiting excess profit resources, they eventually make them vanish.

The fundamental price of a stock reflects the profit that the company is expected to make. If the trade price is below the fundamental price then traders would be eager to buy the stock because they could acquire the profit of the company at less than its value. And vice-versa, they would be eager to sell an overpriced stock. Consequently, if a security trades at a price far from its fundamental value, then it is more likely that the trade price will move towards the fundamental value, which implies predictability and excess profit opportunity. Therefore such an opportunity will be exploited and should finally vanish.

That is why it is conventionally assumed that trading prices are consistent and follow the fundamental prices. This is implied in the so-called efficient market hypothesis [52]. This widely accepted notion is typically classified into three categories:

The weak-form efficiency assumes that no excess returns can be earned by using investment strategies based on historical stock prices or other financial data. This implies that there are no statistically observable patterns which would enable traders to consistently produce excess returns based on quantitative analysis
techniques.

The *semistrong-form efficiency* says that stock prices adjust within an arbitrarily small but finite amount of time in an unbiased fashion to publicly available new information, so that no excess returns can be earned by trading on that information. The semi-strong form of efficiency implies that fundamental analysis techniques will not be able to produce excess returns reliably. The only possibility of making excess profit is through private information.

The *strong-form efficiency* states that stock prices reflect all information and no one can achieve excessive returns as all information is public. A trader can outperform the market only because of pure luck. With no information arriving in the market, prices would not change at all.

Even though most of the academic economic forums assume that at least the weak form of the efficient market hypothesis holds for the real financial markets, there is no clear evidence of this. Furthermore, it is not clear if there is a hierarchy of the weak and semi-strong forms in the practice. The transaction prices are set by trading, which happens much more frequently than news arrival. Stock prices change on a much shorter time scale than the “true values” of the companies could. Variations in price due to imbalances in supply and demand are often greater in magnitude than some of the news would imply. Those extreme price changes and stock market crashes which happen with no news arriving in the market could not be explained within the framework of efficient market hypothesis. To demonstrate this, in Table 1.1 we list the 12 largest price changes in the S&P 500 index of the New York Stock Exchange between 1946 and 1987, and the explanations provided by the *New York Times* [20]. We even find examples when the journal states that there was no reason for a large price movement. If the entire market can drop more than 20% during a day when no major news is occurring, then the stocks must have been mispriced either before the crash or after, or perhaps both. These facts suggest that some of the extreme events may originate in the dynamics of the complex system of trading. Fluctuations should not be neglected, not even in equilibrium. Markets, in general, are not perfectly efficient, and there are time periods when they are quite far from it.

Although the robust process of market impact ensures that markets tend to be close to efficiency, it also explains why perfect efficiency cannot be achieved. There is a driving force to make the market efficient but only if the market is not in an efficient state. When the market is efficient, there is no guarantee it will remain so. Consequently, prices should fluctuate around their efficient values, but not necessarily take them permanently. It is similar to a physical oscillator: just because the resultant force typically points toward the equilibrium point, it does not mean that the oscillator is at rest in the equilibrium state.

Nonetheless, the notion of market efficiency is meaningful and realized to some extent in the real financial markets. It expresses an important principle: future price movements are expected to have minimal or no predictability, which would enable traders to earn extra profit.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Date</th>
<th>%</th>
<th>NY Times explanation</th>
</tr>
</thead>
</table>
| 1    | Oct 19, 1987 | -20.5 | Worry over dollar decline and rate deficit  
|      |             |     | Fear of US not supporting dollar                                                          |
| 2    | Oct 21, 1987 |  9.1 | Interest rates continue to fall  
|      |             |     | Deficit talks in Washington                                                                |
| 3    | Oct 26, 1987 | -8.3 | Fear of budget deficits  
|      |             |     | Margin calls  
|      |             |     | Reaction to falling foreign stocks                                                           |
| 4    | Sept 3, 1946 | -6.7 | “... no basic reason for the assault on prices.”                                     |
| 5    | May 28, 1962 | -6.7 | Kennedy forces rollback of steel price hike                                           |
| 6    | Sept 26, 1955 | -6.6 | Eisenhower suffers heart attack                                                        |
| 7    | Jun 28, 1955 | -5.4 | Outbreak of Korean War                                                             |
| 8    | Oct 20, 1987 |  5.3 | Investors looking for quality stocks                                                   |
| 9    | Sept 9, 1946 | -5.2 | Labor unrest in maritime and trucking                                                 |
| 10   | Oct 16, 1987 | -5.2 | Fear of trade deficit  
|      |             |     | Fear of higher interest rates                                                          |
|      |             |     | Tension with Iran                                                                     |
| 11   | May 27, 1970 |  5.0 | Rumors of change in economic policy                                                   |
|      |             |     | “stock surge happened for no fundamental reason”                                      |
| 12   | Sept 11, 1986 | -4.8 | Foreign governments refuse to lower interest rates                                  |
|      |             |     | Cracking on triple witching announced                                                 |

Table 1.1: The New York Times’ explanations for the 12 largest one day changes in the S&P 500 index between 1946 and 1987 [20]. Each negative change corresponds to a stock market crash. The red lines indicate news that involves relevant market information. Note that the even the largest changes could not be sufficiently explained by news.
1.4.4 Price diffusion

The fact that a market is expected to be almost efficient has a profound consequence on the price formation, which we will discuss in this section.

Let us consider a series of stock transactions and denote the transaction price right after the $i$th trade at time $t_i$ by $P_i$. For the sake of simplicity let us neglect the typical long time-scale increase in the fundamental value of the company and restrict our focus to the fluctuations of the price. In this case, the efficient market hypothesis requires that

$$E[P_i - P_{i-n}] = 0 \quad \forall n > 0$$

at any given time $t_i$, where $E[.]$ refers to the expected value. Otherwise, one would easily make excess profit statistically. However, $P_i$ is not necessarily equal to $P_{i-n}$, thus $\Delta P_i^{(n)} = P_i - P_{i-n}$ price increments can be characterized by a zero mean probability distribution. Hence, $P(t)$ price must follow a one-dimensional random walk path along the price axis. As a consequence, the discipline of price changes, such as scaling properties, is similar to that of a diffusive particle, and that is why the process of price formation is often referred to as price diffusion in the literature.

Note that even the weak form of market efficiency makes stronger assumptions than what Eq. 1.2 implies. If there is no statistically exploitable historical pattern, that is, future prices cannot be predicted,

$$E[P_i|P_1, P_2, \ldots, P_{i-1}] = P_{i-1} \quad \forall i$$

should also hold. A stochastic process abiding by Eq. 1.3 is called a martingale.

The martingale allows for a great variety of random walks. Any process with independent identically distributed finite $\Delta P_i^{(1)}$ increments with zero mean satisfies Eq. 1.3. However, there are many non-independent stochastic processes which are also martingales. Real price time series are not perfect martingales since markets are not perfectly efficient. Furthermore, stock prices are normally expected to increase, reflecting the increase in the fundamental value due to the profit from the production of the company (unless the entire profit is paid off to the investors as a dividend). This may be taken into account by extending Eq. 1.3 as follows

$$E[P_i|P_1, P_2, \ldots, P_{i-1}] = P_{i-1} + \mu(t_i, t_{i-1})$$

where $\mu(t_i, t_{i-1})$ is the expected fundamental increase which depends on the $[t_i, t_{i-1}]$ real-time interval. However, in real markets $|\mu(t_i, t_{i-1})| \ll |\Delta P_i^{(1)}|$ for most of the $i$-s, thus the $\mu$ term may be neglected in an analysis of price increments on short time scales (typically shorter than a couple of months). However, the $\mu(t_i, t_{i-1})$ time series are strongly correlated and may be considered as a drift in the diffusion process, as opposed to $\Delta P_i^{(1)}$ price increments, which are basically uncorrelated. Consequently it may dominate price changes on long enough time scales (i.e., a few years). In mathematical terms, $|\mu(t_i, t_{i-n})| \gg |\Delta P_i^{(n)}|$ for most
of the $i$-s if $n$ is large enough so $t_i - t_{i-n}$ is in the order of magnitude of years. In other words, stock price changes are dominated by changes in fundamental value for long time scales.

### 1.4.5 Stylized facts

Analysis based on historical transaction price time series of financial assets revealed non-trivial statistical regularities of the price diffusion. The three most obtrusive of these properties became commonly known as stylized facts.

Just as in the case of any kind of random walk, the increments on longer time scales can be derived from the elementary increments, that is

$$
\Delta P_i^{(n)} = \sum_{j=0}^{n-1} \Delta P_{i-j}^{(1)}.
$$

According to the central limit theorem, if $\Delta P_i^{(1)}$ time series is well behaved then the $\Delta P_i^{(n)}$ should be well described as an independent, normally distributed stochastic process for any sufficiently large $n$. For example, if increments were independent, uniformly distributed random variables, then the Gaussian description would be a good approximation from $n \approx 10$. This is why Bachelier assumed in his early work that the price changes are well modeled by a Brownian motion in the continuous limit [21]. However, careful data analysis pointed out anomalies in the process of price diffusion. The trade prices seem to exhibit super-diffusion, i.e., the tails of the probability density function of the distribution of price changes decay more slowly than any exponential ones.

Obviously, the origin of the anomaly cannot be rooted in the autocorrelation of the price increment time series. The first stylized fact is the discovery that the autocorrelations are weak and they decay exponentially with a characteristic time of a couple of minutes [23]. This is in accordance with the weak form of the efficient market hypothesis. Indeed, a weak autocorrelation does not necessarily contradict the hypothesis. All the trades have transaction costs and if the predictability is too weak it cannot be exploited profitably.

The second stylized fact is the observation of the fat tails of the distribution of the price changes. In reality, the fat tails are present at almost any time scale (excluding extremely long ones). Mandelbrot [22] suggested that the distribution of price fluctuations has a stable Pareto-Levy form with fat power law tails (with infinite second moment). Just like the Gaussian, Pareto-Levy is a stable distribution under addition, and it is the limit distribution of those with infinite variance. This could explain the persistence of the fat tails for the different time scales. More recently, this observation was refined by others, who found that for exceptionally large fluctuations, the fat tails of Pareto-Levy distribution overestimate the actual probabilities. The current consensus is that for very large fluctuations Pareto-Levy power law tails crossover to either exponential [23] or to a steeper power law [26], which both have finite second moment and ensure that the distribution eventually converges to the Gaussian for large time scales. However, the convergence is very slow as the power law nature of the tails is
stable under addition even for finite second moment. Although the middle of the
distribution is approaching a Gaussian, the tails far out remain power laws. It is
only the crossover points between the Gaussian and the power law regions that
move toward the plus and minus infinity as the time scale increases. Depending
on the exponent of the power law, the Gaussian may not be a good approximation
even at $n \approx 10^5$. Note that the fat tails imply extreme risk, that is, rare large
price changes are likely to be much larger than one would expect for a Gaussian.
This fact is especially important from the aspect of risk management.

The third stylized fact is the time series of magnitude of price changes, i.e.,
the volatility, which shows autocorrelation that decays as a power-law, reflecting
long-term temporal correlations. The scale-free characteristics are observed for
time scales from a couple of minutes all the way up to one year or longer [24].

Note that for real markets $|\Delta P_i^{(n)}| \ll P_{i-n}$ typically, if $n$ is not too large.
Consequently the stylized facts can be observed for the relative price increments
$\Delta P_i^{(n)}/P_{i-n}$ and for the logarithmic returns

$$\Delta r_i^{(n)} = \log P_i - \log P_{i-n} = \log \left(1 + \frac{\Delta P_i^{(n)}}{P_{i-n}}\right) \approx \frac{\Delta P_i^{(n)}}{P_{i-n}}$$

in the same fashion as for $\Delta P_i^{(n)}$ price increments, provided the time scale is not
too long.

The importance of the stylized facts is rooted in their robust presence in
real data. Stylized facts can be observed (1) for different asset types (2) across
different markets, (3) for different time periods, and (4) for different time scales.
The origins of the second and third stylized facts are not yet known, however,
their implications are enormous. The stylized facts are suspected of being a sign
of the existence of a dominant underlying microscopic mechanism in the economy,
which is perhaps rooted in the mechanism of trading.

Physicists may be interested in power-laws observed in economic data because
they might be related to critical behavior. The second and third stylized facts
also involve power law characteristics. However, there is no good evidence of
universality in trading data in the statistical physics sense [25].

1.4.6 Limit order based trading mechanism

We have referred to the elementary trading mechanism many times in the in-
troduction. In this section we describe it in detail using the concept of limit
orders.

The mismatch between buyers and sellers that typically exists at any given
instant in the market is solved via an order-based market with two basic kinds of
orders. Impatient traders submit market orders, which are requests to buy or sell
a given number of shares immediately at the best available price. More patient
traders submit limit orders, or quotes which also state a limit price, corresponding
to the worst allowable price for the transaction. Limit orders often fail to result

\footnote{Note that the word “quote” can be used either to refer to the limit price or to the limit
order itself.}
Figure 1.2: A schematic illustration of the continuous double auction mechanism. Limit orders are stored in the limit order book. We adopt the arbitrary convention that buy orders are negative and sell orders are positive. As a market order arrives, it has transactions with limit orders of the opposite sign, in order of price (first) and time of arrival (second). The best quotes at prices \( a(t) \) or \( b(t) \) move whenever an incoming market order has sufficient size to fully deplete the stored volume at \( a(t) \) or \( b(t) \).

in an immediate transaction, and are stored in a queue called the limit order book. Buy limit orders are called bids, and sell limit orders are called offers or asks. We use the logarithmic price \( a(t) \) to denote the position of the best (lowest) offer and \( b(t) \) for the position of the best (highest) bid. These are also called the inside quotes. There is typically a non-zero price gap between them, called the spread \( s(t) = a(t) - b(t) \). Prices are not continuous, but rather have discrete quanta called ticks. Throughout the rest of the study, all prices will be expressed as logarithms, and to avoid endless repetition, the word price will mean the logarithm of the price. The minimum interval that prices change on is the tick size \( dp \) (also defined on a logarithmic scale; note this is not true for real markets). Note that \( dp \) is not necessarily infinitesimal.

As market orders arrive they are matched against limit orders of the opposite sign in order of first price and then arrival time, as shown in Figure 1.2. Because orders are placed for varying numbers of shares, matching is not necessarily one-to-one. For example, suppose the best offer is for 200 shares at $60 and the next best is for 300 shares at $60.25; a buy market order for 250 shares buys 200 shares at $60 and 50 shares at $60.25, moving the best offer \( a(t) \) from $60 to $60.25. A high density of limit orders per price results in high liquidity for market orders, i.e., it decreases the price movement when a market order is placed. Let \( n(p,t) \) be the stored density of limit order volume at price \( p \), which we will call the depth profile of the limit order book at any given time \( t \). The total stored limit order volume at price level \( p \) is \( n(p,t)dp \). For unit order size the shift in the best ask
$a(t)$ produced by a buy market order is given by solving the equation

$$\omega = \sum_{p=a(t)}^{p^t} n(p, t)dp$$

for $p^t$. The shift in the best ask is $p^t - a(t)$, which is the instantaneous price impact for buy market orders. A similar statement applies for sell market orders, where the price impact can be defined in terms of the shift in the best bid. (Alternatively, it is also possible to define the price impact in terms of the change in the midpoint price.)

We will refer to a buy limit order whose limit price is greater than the best ask, or a sell limit order whose limit price is less than the best bid, as a crossing limit order or marketable limit order. Such limit orders result in immediate transactions, with at least part of the order immediately executed.

The limit-order based trading mechanism is a robust rule of trading. Many modern financial markets list it in their trading guides as an institutional rule, however, it would emerge naturally as it is in the interests of those traders who make the transaction with each other at last. The buyer receives the lowest price available in the market, whereas the seller can sell the assets at the highest possible price at the moment. Although operational rules of different stock markets may seem to differ significantly, these principles of the trading are the same. The limit-order based trading mechanism attains a continuous double auction and it is an ideal, yet realistic procedure of trading. It clarifies how the (unveiled) supply and demand should be defined and the algorithm to match them. It translates trader actions into price formation. From the sequence of events, that is, the series of order placements and cancellations, one can uniquely reconstruct the transactions and the price time series, and derive many related quantities, such as the bid-ask spread, volatility, price impact, liquidity, and price diffusion rate. Since all of these quantities are determined by the same sequence of events, i.e., by the event flow, the mechanism makes possible to unveil relationship between them. The limit-order based trading mechanism also provides the framework for analyzing trader behavior from price formation properties, such as the implications of market efficiency for the order placement.

In the later chapters we will analyze and model stock markets from the aspect of limit order book based trading.
Chapter 2

A simplistic limit order based trading model

2.1 Background

2.1.1 Motivation

In this chapter we analyze the continuous double auction trading mechanism defined in Section 1.4.6 under the assumption of random order flow, developing a model introduced in Daniels et al. [27]1. This analysis produces quantitative predictions about the most basic properties of markets, such as volatility, depth of stored supply and demand, the bid-ask spread, the price impact, and probability and time to fill2 [29]. These predictions are based on the rate at which orders flow into the market, and other parameters of the market, such as order size and tick size. The predictions are falsifiable with no free parameters. This extends the original random walk model of Bachelier [21] by providing a basis for the diffusion rate of prices. The model also provides a possible explanation for the highly concave nature of the price impact function. Even though some of the assumptions of the model are too simple to be literally true, the model provides a foundation onto which more realistic assumptions may easily be added.

The model demonstrates the importance of financial institutions in setting prices, and how solving a necessary economic function such as providing liquidity can have unanticipated side-effects. In a world of imperfect rationality and imperfect information, the task of demand storage necessarily causes persistence. Under perfect rationality all traders would instantly update their orders with the arrival of each piece of new information, but this is clearly not true for real markets. The limit order book, which is the queue used for storing unexecuted orders, has long memory when there are persistent orders. It can be regarded as a device for storing supply and demand, somewhat like a capacitor is a device for storing charge. We show that even under completely random IID order flow, the price process displays anomalous diffusion and interesting temporal structure.

1This paper is a published version of the original preprint [28].
2Note that the term prediction is not used in the sense of time-wise prediction. Rather, it refers to relationships between relevant quantities.
The converse is also interesting: for prices to be effectively random, incoming order flow must be non-random, in just the right way to compensate for the persistence.

This work is also of interest from a fundamental point of view because it suggests an alternative approach to doing economics. The assumption of perfect rationality has been popular in economics because it provides a parsimonious model that makes strong predictions. In the spirit of Gode and Sunder [30], we show that the opposite extreme of zero intelligence random behavior provides another reference model that also makes very strong predictions. Like perfect rationality, zero intelligence is an extreme simplification that is obviously not literally true. But as we show here, it provides a useful tool for probing the behavior of financial institutions. The resulting model may easily be extended by introducing simple boundedly rational behaviors. We also differ from standard treatments in that we do not attempt to understand the properties of prices from fundamental assumptions about utility. Rather, we split the problem in two. We attempt to understand how prices depend on order flow rates, leaving the problem of what determines these order flow rates for the future.

One of our main results concerns the average price impact function. The liquidity for executing a market order can be characterized by a price impact function $\Delta p = \phi(\omega, \tau, t)$. $\Delta p$ is the shift in the logarithm of the price at time $t + \tau$ caused by a market order of size $\omega$ placed at time $t$. Understanding price impact is important for practical reasons such as minimizing transaction costs, and also because it is closely related to an excess demand function, providing a natural starting point for theories of statistical or dynamical properties of markets [31, 32]. A naive argument predicts that the price impact $\phi(\omega)$ should increase at least linearly. This argument is as follows: fractional price changes should not depend on the scale of price. Suppose buying a single share raises the price by a factor $k > 1$. If $k$ is constant, buying $\omega$ shares in succession should raise it by $k^\omega$. Thus, if buying $\omega$ shares all at once affects the price at least as much as buying them one at a time, the ratio of prices before and after impact should increase at least exponentially. Taking logarithms implies that the price impact as we have defined it above should increase at least linearly.\footnote{In financial models it is common to define an excess demand function as demand minus supply; when the context is clear, the modifier “excess” is dropped, so that demand refers to both supply and demand.}

In contrast, from empirical studies $\phi(\omega)$ for buy orders appears to be concave [33, 34, 35, 36, 37, 38]. Lillo et al. have shown for that for stocks in the NYSE the concave behavior of the price impact is quite consistent across different stocks [38]. Our model produces concave price impact functions that are in qualitative agreement with these results.

This work also demonstrates the value of physics techniques for economic problems. The analysis makes extensive use of dimensional analysis. We also developed a solution based on a master equation through a generating func-

\footnote{This has practical implications. It is common practice to break up orders in order to reduce losses due to market impact. With a sufficiently concave market impact function, in contrast, it is cheaper to execute an order all at once.}
Figure 2.1: A schematic illustration of our model. The limit order rate is characterized by $\alpha$ with dimension of shares per unit price and per unit time. The market order arrival parameter $\mu$ is measured in shares per unit time. The model is symmetric in the buy and sell sides.

2.1.2 The model

The basis of this model is formed by the limit order based trading rules as they were defined in Section 1.4.6. This model introduced in reference [27], is designed to be as analytically tractable as possible while capturing key features of the continuous double auction. All the order flows are modeled as Poisson processes. We assume that market orders arrive in chunks of $\sigma$ shares, at a rate of $\mu$ shares per unit time. The market order may be a ‘buy’ order or a ‘sell’ order with equal probability. (Thus, the rate at which buy orders or sell orders arrive individually is $\mu/2$.) Limit orders arrive in chunks of $\sigma$ shares as well, at a rate $\alpha$ shares per unit price and per unit time for buy orders and also for sell orders. Offers are placed with uniform probability at integer multiples of a tick size $dp$ in the range of price $b(t) < p < \infty$, and similarly for bids on $-\infty < p < a(t)$. When a market order arrives, it causes a transaction; under the assumption of constant order size, a buy market order removes an offer at price $a(t)$, and if it was the last offer at that price, moves the best ask up to the next occupied price tick. Similarly, a sell market order removes a bid at price $b(t)$, and if it is the last bid at that price, moves the best bid down to the next occupied price tick. In addition, limit orders may also be removed spontaneously by being canceled or by expiring, even without a transaction having taken place. We model this by letting them be removed randomly with constant probability $\delta$ per unit time.

While the assumption of limit order placement over an infinite interval is clearly unrealistic, it provides a tractable boundary condition for modeling the behavior of the limit order book near the midpoint price $m(t) = (a(t) + b(t))/2$,
which is the region of interest since it is where transactions occur. Limit orders far from the midpoint are usually canceled before they are executed (we demonstrate this later in Figure 2.5), and so far from the midpoint, limit order arrival and cancellation have a steady state behavior characterized by a simple Poisson distribution. Although under the limit order placement process the total number of orders placed per unit time is infinite, the order placement per unit price interval is bounded and thus the assumption of an infinite interval creates no problems. Indeed, it guarantees that there are always an infinite number of limit orders of both signs stored in the book, so that the bid and ask are always well-defined and the book never empties. (Under other assumptions about limit order placement this is not necessarily true, as we later demonstrate in Figure A.16.)

In this model, to keep things simple, we are using the conceptual simplification of effective market orders and effective limit orders. When a crossing limit order is placed, part of it may be executed immediately. The effect of this part on the price is indistinguishable from that of a market order of the same size. Similarly, given that this market order has been placed, the remaining part is equivalent to a non-crossing limit order of the same size. Thus, a crossing limit order can be modeled as an effective market order followed by an effective (non-crossing) limit order. Working in terms of effective market and limit orders affects data analysis: the effective market order arrival rate $\mu$ combines both pure market orders and the immediately executed components of crossing limit orders, and similarly the limit order arrival rate $\alpha$ corresponds only to the components of limit orders that are not executed immediately. This is consistent with the boundary conditions for the order placement process, since an offer with $p \leq b(t)$ or a bid with $p \geq a(t)$ would result in an immediate transaction, and thus would be effectively the same as a market order. Defining the order placement process with these boundary conditions realistically allows limit orders to be placed anywhere inside the spread.

Another simplification of this model is the use of logarithmic prices, both for the order placement process and for the tick size $dp$. This has the important advantage that it ensures that prices are always positive. In real markets price ticks are linear, and the use of logarithmic price ticks is an approximation that makes both the calculations and the simulation more convenient. We find that the limit $dp \to 0$, where tick size is irrelevant, is a good approximation for many purposes. We find that tick size is less important than other parameters of the problem, which provides some justification for the approximation of logarithmic price ticks.

Assuming a constant probability for cancellation is clearly ad hoc, but in simulations we find that other assumptions with well-defined timescales, such as constant duration time, give similar results. For our analytic model we use a constant order size $\sigma$. In simulations we also use variable order size, e.g., half-normal distributions with standard deviation $\sqrt{\pi/2}\sigma$, which ensures that the mean value remains $\sigma$. As long as these distributions have thin tails, the differences do not qualitatively affect most of the results reported here, except in

---

5 In assigning independently random distributions for the two events, our model neglects the correlation between market and limit order arrival induced by crossing limit orders.
Even though this model is simply defined, the time evolution is not trivial. One can think of the dynamics as being composed of three parts: (1) the buy market order/sell limit order interaction, which determines the best ask; (2) the sell market order/buy limit order interaction, which determines the best bid; and (3) the random cancellation process. Processes (1) and (2) determine each others’ boundary conditions. That is, process (1) determines the best ask, which sets the boundary condition for limit order placement in process (2), and process (2) determines the best bid, which determines the boundary conditions for limit order placement in process (1). Thus processes (1) and (2) are strongly coupled. It is this coupling that causes the bid and ask to remain close to each other, and guarantees that the spread \( s(t) = a(t) - b(t) \) is a stationary random variable, even though the bid and ask are not. It is the coupling of these processes through their boundary conditions that provides the nonlinear feedback that makes the price process complex.

2.1.3 Summary of prior work

There are two independent lines of prior work, one in the financial economics literature, and the other in the physics literature. The models in the economics literature are directed toward empirical analysis, and treat the order process as static. In contrast, the models in the physics literature are conceptual toy models, but they allow the order process to react to changes in prices, and are thus fully dynamic. Our model bridges this gap. This is explained in more detail below.

The first model of this type that we are aware of was due to Mendelson [39], who modeled random order placement with periodic clearing. This was developed along different directions by Cohen et al. [40], who used techniques from queuing theory, but assumed only one price level and addressed the issue of time priority at that level (motivated by the existence of a specialist who effectively pinned prices to make them stationary). Domowitz and Wang [41] and Bollerslev et al. [42] further developed this to allow more general order placement processes that depend on prices, but without solving the full dynamical problem. This allows them to get a stationary solution for prices. In contrast, in our model the prices that emerge make a random walk, and so are much more realistic. In order to get a solution for the depth of the order book we have to go into price coordinates that co-move with the random walk. Dealing with the feedback between order placement and prices makes the problem much more difficult, but it is key for getting reasonable results.

The models in the physics literature incorporate price dynamics, but have tended to be conceptual toy models designed to understand the anomalous diffusion properties of prices. This line of work begins with a paper by Bak et al. [43] which was developed by Eliezir and Kogan [44] and by Tang [45]. They assume that limit orders are placed at a fixed distance from the midpoint, and that the limit prices of these orders are then randomly shuffled until they result in transactions. It is the random shuffling that causes price diffusion. This assumption, which we feel is unrealistic, was made to take advantage of the analogy to a stan-
standard reaction-diffusion model in the physics literature. Maslov [46] introduced an alternative model that was solved analytically in the mean-field limit by Slanina [47]. Each order is randomly chosen to be either a buy or a sell, and either a limit order or a market order. If a limit order, it is randomly placed within a fixed distance of the current price. This again gives rise to anomalous price diffusion. A model allowing limit orders with Poisson order cancellation was proposed by Challet and Stinchcombe [48]. Iori and Chiarella [49] have numerically studied a model including fundamentalists and technical traders.

The model studied in this chapter was introduced by Daniels et al. [27]. This adds to the literature by introducing a model that treats the feedback between order placement and price movement, while having enough realism so that the parameters can be tested against real data. The prior models in the physics literature have tended to focus primarily on the anomalous diffusion of prices. While interesting and important for refining risk calculations, this is a second-order effect. In contrast, we focus on the first order effects of primary interest to market participants, such as the bid-ask spread, volatility, depth profile, price impact, and the probability and time to fill an order. We demonstrate how dimensional analysis becomes a useful tool in an economic setting, and develop mean field theories in a context that is more challenging than that of the toy models of previous work.

Subsequent to work by Daniels et al. [27], Bouchaud et al. [50] demonstrated that, under the assumption that prices execute a random walk, by introducing an additional free parameter they can derive a simple equation for the depth profile. In this study we show how to do this from first principles without introducing a free parameter.

2.2 Predictions of the model

In this section we give an overview of the phenomenology of the model. Because this model has five parameters, understanding all their effects would generally be a complicated problem in and of itself. This task is greatly simplified by the use of dimensional analysis, which reduces the number of independent parameters from five to two. Thus, before we can even review the results, we need to first explain how dimensional analysis applies in this setting. One of the surprising aspects of this model is that one can derive several powerful results using the simple technique of dimensional analysis alone.

Unless otherwise mentioned, the results presented in this section are based on simulations. These results are compared to theoretical predictions in Appendix A.

2.2.1 Dimensional analysis

For this problem the three fundamental dimensions in the model are shares, price, and time. Note that by price, we mean the logarithm of price; as long as we are consistent, this does not create problems with the dimensional analysis. There are five parameters: three rate constants and two discreteness parameters. The order flow rates are $\mu$, the market order arrival rate, with dimensions of shares
per time; $\alpha$, the limit order arrival rate per unit price, with dimensions of $\text{shares per price per time}$; and $\delta$, the rate of limit order decays, with dimensions of $1$/time. These play a role similar to rate constants in physical problems. The two discreteness parameters are the price tick size $dp$, with dimensions of $\text{price}$, and the order size $\sigma$, with dimensions of $\text{shares}$. This is summarized in Table 2.1.

Dimensional analysis can be used to reduce the number of relevant parameters. Because there are five parameters and three dimensions ($\text{price}$, $\text{shares}$, $\text{time}$), and because in this case the dimensionality of the parameters is sufficiently rich, the dimensional relationships reduce the degrees of freedom, so that all the properties of the limit-order book can be described by functions of two parameters. It is useful to construct these two parameters so that they are nondimensional.

We perform the dimensional reduction of the model by guessing that the effect of the order flow rates is primary to that of the discreteness parameters. This leads us to construct nondimensional units based on the order flow parameters alone, and take nondimensionalized versions of the discreteness parameters as the independent parameters whose effects remain to be understood. As we will see, this is justified by the fact that many of the properties of the model depend only weakly on the discreteness parameters. We can thus understand much of the richness of the phenomenology of the model through dimensional analysis alone.

There are three order flow rates and three fundamental dimensions. If we temporarily ignore the discreteness parameters, there are unique combinations of the order flow rates with units of shares, price, and time. These define a characteristic number of shares $N_c = \mu/2\delta$, a characteristic price interval $p_c = \mu/2\alpha$, and a characteristic timescale $t_c = 1/\delta$. This is summarized in Table 2.2. The factors of two occur because we have defined the market order rate for either a buy or a sell order to be $\mu/2$. We can thus express everything in the model in nondimensional terms by dividing by $N_c$, $p_c$, or $t_c$ as appropriate, e.g., to measure shares in nondimensional units $\hat{N} = N/N_c$, or to measure price in nondimensional units $\hat{p} = p/p_c$.

The value of using nondimensional units is illustrated in Figure 2.2. Figure 2.2 (a) shows the average depth profile for three different values of $\mu$ and $\delta$ with the other parameters held fixed. When we plot these results in dimensional units, the results look quite different. However, when we plot them in terms of nondimensional units, as shown in Figure 2.2(b), the results are indistinguishable. As explained below, because we have kept the nondimensional order size fixed,
Table 2.2: Important characteristic scales and nondimensional quantities. We summarize the characteristic share size, price, and times defined by the order flow rates, as well as the two nondimensional scale parameters $dp/p_c$ and $\epsilon$ that characterize the effect of finite tick size and order size. Dimensional analysis makes it clear that all the properties of the limit order book can be characterized in terms of functions of these two parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_c$</td>
<td>characteristic number of shares</td>
<td>$\mu/2\delta$</td>
</tr>
<tr>
<td>$p_c$</td>
<td>characteristic price interval</td>
<td>$\mu/2\alpha$</td>
</tr>
<tr>
<td>$t_c$</td>
<td>characteristic time</td>
<td>$1/\delta$</td>
</tr>
<tr>
<td>$dp/p_c$</td>
<td>nondimensional tick size</td>
<td>$2\alpha dp/\mu$</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>nondimensional order size</td>
<td>$2\delta\sigma/\mu$</td>
</tr>
</tbody>
</table>

Figure 2.2: The usefulness of nondimensional units. (a) We show the average depth profile for three different parameter sets. The parameters $\alpha = 0.5$, $\sigma = 1$, and $dp = 0$ are held constant, while $\delta$ and $\mu$ are varied. The line types are: (dotted) $\delta = 0.001$, $\mu = 0.2$; (dashed) $\delta = 0.002$, $\mu = 0.4$ and (solid) $\delta = 0.004$, $\mu = 0.8$. (b) is the same, but plotted in nondimensional units. The horizontal axis has units of price, and so has nondimensional units $\hat{p} = p/p_c = 2\alpha p/\mu$. The vertical axis has units of $n$ shares/price, and so has nondimensional units $\hat{n} = np_c/N_c = n\delta/\alpha$. Because we have chosen the parameters to keep the nondimensional order size $\epsilon$ constant, the collapse is perfect. Varying the tick size has little effect on the results other than making them discrete.
the collapse is perfect. Thus, the problem of understanding the behavior of this model is reduced to studying the effect of tick size and order size.

To understand the effect of tick size and order size it is useful to do so in nondimensional terms. The nondimensional scale parameter based on tick size is constructed by dividing by the characteristic price, i.e., \( dp/p_c = 2\alpha dp/\mu \). The theoretical analysis and the simulations show that there is a sensible continuum limit as the tick size \( dp \to 0 \), in the sense that there is non-zero price diffusion and a finite spread. Furthermore, the dependence on tick size is weak, and for many purposes the limit \( dp \to 0 \) approximates the case of finite tick size fairly well. As we will see, working in this limit is essential for getting tractable analytic results.

A nondimensional scale parameter based on order size is constructed by dividing the typical order size (which is measured in shares) by the characteristic number of shares \( N_c \), i.e., \( \epsilon \equiv \sigma/N_c = 2\delta \sigma/\mu \). \( \epsilon \) characterizes the “chunkiness” of the orders stored in the limit order book. As we will see, \( \epsilon \) is an important determinant of liquidity, and it is a particularly important determinant of volatility. In the continuum limit \( \epsilon \to 0 \) there is no price diffusion. This is because price diffusion can occur only if there is a finite probability for price levels outside the spread to be empty, thus allowing the best bid or ask to make a persistent shift. If we let \( \epsilon \to 0 \) while the average depth is held fixed, the number of individual orders becomes infinite, and the probability that spontaneous decays or market orders can create gaps outside the spread becomes zero. This is verified in simulations. Thus, the limit \( \epsilon \to 0 \) is always a poor approximation of a real market. \( \epsilon \) is a more important parameter than the tick size \( dp/p_c \). In the mean field analysis in Appendix A, we let \( dp/p_c \to 0 \), reducing the number of independent parameters from two to one, and in many cases find that this is a good approximation.

The order size \( \sigma \) can be thought of as the order granularity. Just as the properties of a beach with fine sand are quite different from that of one populated by fist-sized boulders, a market with many small orders behaves quite differently from one with a few large orders. \( N_c \) provides the scale against which the order size is measured, and \( \epsilon \) characterizes the granularity in relative terms. Alternatively, \( 1/\epsilon \) can be thought of as the annihilation rate from market orders expressed in units of the size of spontaneous decays. Note that in nondimensional units the number of shares can also be written \( \hat{N} = N/N_c = N\epsilon/\sigma \).

The construction of the nondimensional granularity parameter illustrates the importance of including a spontaneous decay process in this model. If \( \delta = 0 \) (which implies \( \epsilon = 0 \)), there is no spontaneous decay of orders, and depending on the relative values of \( \mu \) and \( \alpha \), generically either the depth of orders will accumulate without bound or the spread will become infinite. As long as \( \delta > 0 \), in contrast, this is not a problem.

For some purposes the effects of varying tick size and order size are fairly small, and we can derive approximate formulas using dimensional analysis based only on the order flow rates. For example, in Table 2.3 we give dimensional scaling formulas for the average spread, the market order liquidity (as measured by the average slope of the depth profile near the midpoint), the volatility, and the asymptotic depth (defined below). Because these estimates neglect the effects
Table 2.3: Estimates from dimensional analysis for the scaling of a few market properties based on order flow rates alone. $\alpha$ is the limit order density rate, $\mu$ is the market order rate, and $\delta$ is the spontaneous limit order removal rate. These estimates are constructed by taking the combinations of these three rates that have the proper units. They neglect the dependence on the order granularity $\epsilon$ and the nondimensional tick size $dp/p_c$. More accurate relations from simulation and theory are given in Table 2.4.

An approximate formula for the mean spread can be derived by noting that it has dimensions of price, and the unique combination of order flow rates with these dimensions is $\mu/\alpha$. While the dimensions indicate the scaling of the spread, they cannot determine multiplicative factors of order unity. A more intuitive argument can be made by noting that inside the spread, removal due to cancellation is dominated by removal due to market orders. Thus, the total limit order placement rate inside the spread, for either buy or sell limit orders $\alpha s$, must equal the order removal rate $\mu/2$, which implies that spread is $s = \mu/2\alpha$. As we will see later, this argument can be generalized and made more precise within our mean-field analysis, which then also predicts the observed dependence on the granularity parameter $\epsilon$. However, this dependence is rather weak and only causes a variation of roughly a factor of two for $\epsilon < 1$ (see Figures 2.10 and A.10), and the factor of 1/2 derived above is a good first approximation. Note that this prediction of the mean spread is just the characteristic price $p_c$.

It is also easy to derive the mean asymptotic depth, which is the density of shares far away from the midpoint. The asymptotic depth is an artificial construct of our assumption of order placement over an infinite interval; it should be regarded as providing a simple boundary condition so that we can study the behavior near the midpoint price. The mean asymptotic depth has dimensions of $shares/price$, and is, therefore, given by $\alpha/\delta$. Furthermore, because removal by market orders is insignificant in this regime, it is determined by the balance between order placement and decay, and far from the midpoint the depth at any given price is Poisson distributed. This result is exact.

The average slope of the depth profile near the midpoint is an important
Table 2.4: The dependence of market properties on model parameters based on simulation and theory, with the relevant figure numbers. These formulas include corrections for order granularity $\epsilon$ and finite tick size $dp/p_c$. The formula for asymptotic depth from dimensional analysis in Table 2.3 is exact with zero tick size. The expression for the mean spread is modified by a function of $\epsilon$ and $dp/p_c$, though the dependence on them is fairly weak. For the liquidity $\lambda$, corresponding to the slope of the depth profile near the origin, the dimensional estimate must be modified because the depth profile is no longer linear (mainly depending on $\epsilon$) and so the slope depends on price. The formulas for the volatility are empirical estimates from simulations. The dimensional estimate for the volatility from Table 2.3 is modified by a factor of $\epsilon^{-0.5}$ for the early time price diffusion rate and a factor of $\epsilon^{0.5}$ for the late time price diffusion rate.

determinant of liquidity, since it affects the expected price response when a market order arrives. The slope has dimensions of shares/price$^2$, which implies that in terms of the order flow rates it scales roughly as $\alpha^2/\mu \delta$. This is also the ratio of the asymptotic depth to the spread. As we will see later, this is a good approximation when $\epsilon \sim 0.01$, but for smaller values of $\epsilon$, the depth profile is not linear near the midpoint, and this approximation fails.

The last two entries in Table 2.4 are empirical estimates for the price diffusion rate $D$, which is proportional to the square of the volatility. That is, for normal diffusion, starting from a point at $t = 0$, the variance $\nu$ after time $t$ is $\nu = Dt$. The volatility at any given timescale $t$ is the square root of the variance at timescale $t$. The estimate for the diffusion rate based on dimensional analysis in terms of the order flow rates alone is $\mu^2 \delta/\alpha^2$. However, simulations show that short time diffusion is much faster than long time diffusion, due to negative autocorrelations in the price process, as shown in Figure 2.11. The initial and the asymptotic diffusion rates appear to obey the scaling relationships given in Table 2.4. Though our mean-field theory is not able to predict this functional form, the fact that early and late time diffusion rates are different can be understood within the framework of our analysis, as described in Sec. A.5. Anomalous diffusion of this type implies negative autocorrelations in midpoint prices. Note that we use the term “anomalous diffusion” to imply that the diffusion rate is different on short and long timescales. We do not use this term in the strict sense, i.e., that the long-time diffusion is proportional to $t^\gamma$ with $\gamma \neq 1$ (for long times $\gamma = 1$ in our case).
2.2.2 Varying the granularity parameter $\epsilon$

We first investigate the effect of varying the order granularity $\epsilon$ in the limit $dp \to 0$. As we will see, the granularity has an important effect on most of the properties of the model, and particularly on depth, price impact, and price diffusion. The behavior can be divided into three regimes, roughly as follows:

- **Large $\epsilon$, i.e., $\epsilon \gtrsim 0.1$.** This corresponds to a large accumulation of orders at the best bid and ask, nearly linear market impact, and roughly equal short and long time price diffusion rates. This is the regime where the mean-field approximation used in the theoretical analysis works best.

- **Medium $\epsilon$ i.e., $\epsilon \sim 0.01$.** In this range the accumulation of orders at the best bid and ask is small and near the midpoint price the depth profile increases nearly linearly with price. As a result, as a crude approximation the price impact increases as roughly the square root of order size.

- **Small $\epsilon$ i.e., $\epsilon \lesssim 0.001$.** The accumulation of orders at the best bid and ask is very small, and near the midpoint the depth profile is a convex function of price. The price impact is very concave. The short time price diffusion rate is much greater than the long time price diffusion rate.

Since the results for bids are symmetric with those for offers about $p = 0$, for convenience we only show the results for offers, i.e., buy market orders and sell limit orders. In this subsection, prices are measured relative to the midpoint, and simulations are in the continuum limit where the tick size $dp \to 0$. The results in this section are from numerical simulations. Also, bear in mind that far from the midpoint the predictions of this model are not valid due to the unrealistic assumption of an order placement process with an infinite domain. Thus, the results are potentially relevant to real markets only when the price $p$ is at most a few times as large as the characteristic price $p_c$.

**Depth profile**

The *mean depth profile*, i.e., the average number of shares per price interval, and the mean cumulative depth profile are shown in Figure 2.3, and the standard deviation of the cumulative profile is shown in Figure 2.4. Since the depth profile has units of *shares/price*, nondimensional units of depth profile are \( \hat{n} = np_c/N_c = n\delta/\alpha \). The cumulative depth profile at any given time $t$ is defined as

\[
N(p, t) = \sum_{\tilde{p}=0}^{p} n(\tilde{p}, t)dp.
\]

This has units of shares and so in nondimensional terms is \( \hat{N}(p) = N(p)/N_c = 2\delta N(p)/\mu = N(p)e/\sigma \).

In the high $\epsilon$ regime the annihilation rate due to market orders is low (relative to $\delta \sigma$), and there is a significant accumulation of orders at the best ask, so that the average depth is much greater than zero at the midpoint. The mean depth
Figure 2.3: The mean depth profile and cumulative depth versus $\hat{p} = p/p_c = 2\alpha p/\mu$. The origin $p/p_c = 0$ corresponds to the midpoint. (a) is the average depth profile $n$ in nondimensional coordinates $\hat{n} = np_c/N_c = n\delta/\alpha$. (b) is nondimensional cumulative depth $N(p)/N_c$. We show three different values of the nondimensional granularity parameter: $\epsilon = 0.2$ (solid), $\epsilon = 0.02$ (dash), $\epsilon = 0.002$ (dot), all with tick size $dp = 0$.

profile is a concave function of price. In the medium $\epsilon$ regime the market order removal rate increases, depleting the average depth near the best ask, and the profile is nearly linear over the range $p/p_c \leq 1$. In the small $\epsilon$ regime the market order removal rate increases even further, making the average depth near the ask very close to zero, and the profile is a convex function over the range $p/p_c \leq 1$.

The standard deviation of the depth profile is shown in Figure 2.4. We see that the standard deviation of the cumulative depth is comparable to the mean depth, and that as $\epsilon$ increases, near the midpoint there is a similar transition from convex to concave behavior.

The uniform order placement process seems at first glance one of the most unrealistic assumptions of our model, leading to depth profiles with a finite asymptotic depth (which also implies that there is an infinite number of orders in the book). However, orders far away from the spread in the asymptotic region almost never get executed and thus do not affect the market dynamics. To demonstrate this in Figure 2.5 we show the comparison between the limit-order depth profile and the depth $n_e$ of only those orders which eventually get executed.\(^6\) The density $n_e$ of executed orders decreases rapidly as a function of the distance from the mid-price. Therefore, we expect that near the midpoint our results should be similar to alternative order placement processes, as long as they also lead to an exponentially decaying profile of executed orders (which is what we observe above). However, to understand the behavior further away from the midpoint it is inevitable that we make enhancements that include more realistic order place-

\(^6\)Note that the ratio $n_e/n$ is not the same as the probability of filling orders (Figure 2.12) because in that case the price $p/p_c$ refers to the distance of the order from the midpoint at the time when it was placed.
Figure 2.4: Standard deviation of the nondimensionalized cumulative depth versus nondimensional price, corresponding to Figure 2.3.

Figure 2.5: A comparison between the depth profiles and the effective depth profiles as defined in the text, for different values of $\epsilon$. Heavy lines refer to the effective depth profiles $n_e$ and the light lines correspond to the depth profiles.
ment processes grounded on empirical measurements of market data.

Liquidity for market orders: the price impact function

In this subsection we study the *instantaneous price impact* function $\phi(t, \omega, \tau \to 0)$. This is defined as the (logarithm of the) midpoint price shift immediately after the arrival of a market order in the absence of any other events. This should be distinguished from the asymptotic price impact $\phi(t, \omega, \tau \to \infty)$, which describes the permanent price shift. While the permanent price shift is clearly very important, we do not study it here. The reader should bear in mind that all prices $p$, $a(t)$, etc. are logarithmic.

The price impact function provides a measure of the liquidity for executing market orders. (The liquidity for limit orders, in contrast, is given by the probability of execution, studied in section 2.2.2). At any given time $t$, the instantaneous ($\tau = 0$) price impact function is the inverse of the cumulative depth profile. This follows immediately from equations (1.4) and (2.1), which in the limit $dp \to 0$ can be replaced by the continuum transaction equation:

$$\omega = N(p, t) = \int_0^p n(\tilde{p}, t) d\tilde{p}. \quad (2.2)$$

This equation makes it clear that at any fixed $t$ the price impact can be regarded as the inverse of the cumulative depth profile $N(p, t)$. When the fluctuations are sufficiently small, we can replace $n(p, t)$ by its mean value $n(p) = \langle n(p, t) \rangle$. In general, however, the fluctuations can be large, and the average of the inverse is not equal to the inverse of the average. There are corrections based on higher order moments of the depth profile, as given in the moment expansion derived in the appendix of the reference [29]. Nonetheless, the inverse of the mean cumulative depth provides a qualitative approximation that gives insight into the behavior of the price impact function. (Note that everything becomes much simpler using medians, since the median of the cumulative price impact function is exactly the inverse of the median price impact.)

Mean price impact functions are shown in Figure 2.6 and the standard deviation of the price impact is shown in Figure 2.7. The price impact exhibits very large fluctuations for all values of $\epsilon$: the standard deviation has the same order of magnitude as the mean or even greater for small $N\epsilon/\sigma$ values. Note that these are actually *virtual price impact* functions. That is, to explore the behavior of the instantaneous price impact for a wide range of order sizes, we periodically compute the price impact that an order of a given size would have caused at that instant, if it had been submitted. We have checked that real price impact curves are the same, but they require a much longer time to accumulate reasonable statistics.

One of the interesting results in Figure 2.6 is the scale of the price impact. The price impact is measured relative to the characteristic price scale $pc$, which as we have mentioned earlier, is roughly equal to the mean spread. As we will argue in relation to Figure 2.8, the range of nondimensional shares shown on the horizontal axis spans the range of reasonable order sizes. This figure demonstrates
Figure 2.6: The average price impact corresponding to the results in Figure (2.3). The average instantaneous movement of the nondimensional mid-price, $\langle dm \rangle / p_c$ caused by an order of size $N/N_c = N\epsilon/\sigma$. $\epsilon = 0.2$ (solid), $\epsilon = 0.02$ (dash), $\epsilon = 0.002$ (dot).

Figure 2.7: The standard deviation of the instantaneous price impact $dm/p_c$ corresponding to the means in Figure 2.6, as a function of normalized order size $\epsilon N/\sigma$. $\epsilon = 0.2$ (solid), $\epsilon = 0.02$ (dash), $\epsilon = 0.002$ (dot).
Figure 2.8: Derivative of the nondimensional mean mid-price movement, with respect to the logarithm of the nondimensional order size $N/N_c = N\epsilon/\sigma$, obtained from the price impact curves in Figure 2.6.

that throughout this range the price is the order of magnitude (and typically less than) the mean spread size.

Due to the accumulation of orders at the ask in the large $\epsilon$ regime, for small $p$ the mean price impact is roughly linear. This follows from equation (2.2) under the assumption that $n(p)$ is constant. In the medium $\epsilon$ regime, under the assumption that the variance in depth can be neglected, the mean price impact should increase as roughly $\omega^{1/2}$. This follows from equation (2.2) under the assumption that $n(p)$ is linearly increasing and $n(0) \approx 0$. (Note that we see this as a crude approximation, but there can be substantial corrections caused by the variance of the depth profile.) Finally, in the small $\epsilon$ regime the price impact is highly concave, increasing much slower than $\omega^{1/2}$. This follows because $n(0) \approx 0$ and the depth profile $n(p)$ is convex.

To get a better feel for the functional form of the price impact function, in Figure 2.8 we numerically differentiate it versus log order size, and plot the result as a function of the appropriately scaled order size. (Note that because our prices are logarithmic, the vertical axis already incorporates the logarithm.) If we were to fit a local power law approximation to the function at each price, this corresponds to the exponent of that power law near that price. Notice that the exponent is almost always less than one, so that the price impact is almost always concave. Making the assumption that the effect of the variance of the depth is not too large, so that equation (2.2) is a good assumption, the behavior of this figure can be understood as follows: for $N/N_c \approx 0$ the price impact is
dominated by \( n(0) \) (the constant term in the average depth profile) and so the logarithmic slope of the price impact is always near to one. As \( N/N_c \) increases, the logarithmic slope is driven by the shape of the average depth profile, which is linear or convex for smaller \( \epsilon \), resulting in concave price impact. For large values of \( N/N_c \), we reach the asymptotic region where the depth profile is flat (and where our model is invalid by design). Of course, there can be deviations to this behavior caused by the fact that the mean of the inverse depth profile is not in general the inverse of the mean, i.e., \( \langle N^{-1}(p) \rangle \neq \langle N(p) \rangle^{-1} \) (see the appendix of [29]).

To compare to real data, note that \( N/N_c = N\epsilon/\sigma \). \( N/\sigma \) is just the order size in shares in relation to the average order size, so by definition it has a typical value of one. For the London Stock Exchange, we have found that typical values of \( \epsilon \) are in the range 0.001 – 0.1. For a typical range of order sizes from 100 – 100,000 shares, with an average size of 10,000 shares, the meaningful range for \( N/N_c \) is, therefore, roughly \( 10^{-5} \) to 1. In this range, for small values of \( \epsilon \) the exponent can reach values as low as 0.2. This offers a possible explanation for the previously mysterious concave nature of the price impact function, and contradicts the linear increase in price impact based on the naive argument presented in the introduction.

**Spread**

The probability density of the spread is shown in Figure 2.9. This shows that the probability density is substantial at \( s/p_c = 0 \). (Remember that this is in the limit \( dp \to 0 \).) The probability density reaches a maximum at a value of the spread approximately 0.2\( p_c \), and then decays. It might seem surprising at first that it decays more slowly for large \( \epsilon \), where there is a large accumulation of orders at the ask. However, it should be borne in mind that the characteristic price
Figure 2.10: The mean value of the spread in nondimensional units $\hat{s} = s/p_c$ as a function of $\epsilon$. This demonstrates that the spread only depends weakly on $\epsilon$, indicating that the prediction from dimensional analysis given in Table 2.3 is a reasonable approximation.

$p_c = \mu/\alpha$ depends on $\epsilon$. Since $\epsilon = 2\delta\sigma/\mu$, by eliminating $\mu$ this can be written $p_c = 2\sigma\delta/(\alpha\epsilon)$. Thus, holding the other parameters fixed, large $\epsilon$ corresponds to small $p_c$, and vice-versa. So in fact, the spread is very small for large $\epsilon$, and large for small $\epsilon$, as expected. The figure just shows the small corrections to the large effects predicted by the dimensional scaling relations.

For large $\epsilon$ the probability density of the spread decays roughly exponentially moving away from the midpoint. This is because for large $\epsilon$ the fluctuations around the mean depth are roughly independent. Thus, the probability for a market order to penetrate to a given price level is roughly the probability that all the ticks smaller than this price level contain no orders, which gives rise to an exponential decay. This is no longer true for small $\epsilon$. Note that for small $\epsilon$ the probability distribution of the spread becomes insensitive to $\epsilon$, i.e. the nondimensionalized distribution for $\epsilon = 0.02$ is nearly the same as that for $\epsilon = 0.002$.

It is apparent from Figure 2.9 that in nondimensional units the mean spread increases with $\epsilon$. This is confirmed in Figure 2.10, which displays the mean value of the spread as a function of $\epsilon$. The mean spread increases monotonically with $\epsilon$. It depends on $\epsilon$ as roughly a constant (equal to approximately 0.45 in nondimensional coordinates) plus a linear term whose slope is rather small. We believe that for most financial instruments $\epsilon < 0.3$. Thus, the variation in the spread caused by varying $\epsilon$ in the range $0 < \epsilon < 0.3$ is not large, and the dimensional analysis based only on rate parameters given in table 2.4 is a good approximation. We get an accurate prediction of the $\epsilon$ dependence across the
Figure 2.11: The variance of the change in the nondimensionalized midpoint price versus the nondimensional time delay interval $\tau \delta$. For a pure random walk this would be a straight line whose slope is the diffusion rate, which is proportional to the square of the volatility. The fact that the slope is steeper for short times comes from the nontrivial temporal persistence of the order book. The three cases correspond to Figure 2.3: $\epsilon = 0.2$ (solid), $\epsilon = 0.02$ (dash), $\epsilon = 0.002$ (dot).

full range of $\epsilon$ from the Independent Interval Approximation technique derived in Section A.7, as shown in Figure A.10.

Volatility and price diffusion

The price diffusion rate, which is proportional to the square of the volatility, is important for determining risk and is a property of central interest. From dimensional analysis in terms of the order flow rates the price diffusion rate has units of $\text{price}^2/\text{time}$, and so must scale as $\mu^2 \delta / \alpha^2$. We can also make a crude argument for this as follows: the dimensional estimate of the spread (see Table 2.4) is $\mu / 2\alpha$. Let this be the characteristic step size of a random walk, and let the step frequency be the characteristic time $1/\delta$ (which is the average lifetime for a share to be canceled). This argument also gives the above estimate for the diffusion rate. However, this is not correct in the presence of negative autocorrelations in the step sizes. The numerical results make it clear that there are important $\epsilon$-dependent corrections to this result, as demonstrated below.

In Figure 2.11 we plot simulation results for the variance of the change in the midpoint price at timescale $\tau$, $\text{Var}(m(t+\tau) - m(t))$. The slope is the diffusion rate, which at any fixed timescale is proportional to the square of the volatility. It appears that there are at least two timescales involved, with a faster diffusion rate for short timescales and a slower diffusion rate for long timescales. Such
anomalous diffusion is not predicted by mean-field analysis. Simulation results show that the diffusion rate is correctly described by the product of the estimate from dimensional analysis based on order flow parameters alone, \( \mu^2 \delta/\alpha^2 \), and a \( \tau \)-dependent power of the nondimensional granularity parameter \( \epsilon = 2 \delta \sigma/\mu \), as summarized in Table 2.4. We cannot currently explain why this power is \(-1/2\) for short-term diffusion and \(1/2\) for long-term diffusion. However, a qualitative understanding can be gained based on the conservation law we derive in Section A.3 in the appendix. A discussion of how this relates to price diffusion is given in Section A.5.

Note that the temporal structure in the diffusion process also implies non-zero autocorrelations of the midpoint price \( m(t) \). This corresponds to weak negative autocorrelations in price differences \( m(t) - m(t-1) \) that persist for timescales until the variance vs. \( \tau \) becomes a straight line. The timescale depends on parameters, but is typically the order of 50 market order arrival times. This temporal structure implies that there exists an arbitrage opportunity which, when exploited, would make prices more random and the structure of the order flow non-random.

**Liquidity for limit orders: probability and time to fill.**

The liquidity for limit orders depends on the probability that they will be filled, and the time to be filled. This obviously depends on price: limit orders close to the current transaction prices are more likely to be filled quickly, while those far away have a lower likelihood of being filled. Figure 2.12 plots the probability \( \Gamma \) of a limit order being filled versus the nondimensionalized price at which it was placed (as with all the figures in this section, this is shown in the midpoint-price centered frame). Figure 2.12 shows that in nondimensional coordinates the probability of filling close to the bid for sell limit orders (or the ask for buy limit orders) decreases as \( \epsilon \) increases. For large \( \epsilon \), this is less than 1 even for negative prices. This says that even for sell orders that are placed close to the best bid there is a significant chance that the offer will be deleted before being executed. This is not true for smaller values of \( \epsilon \), where \( \Gamma(0) \approx 1 \). Far away from the spread the fill probabilities as a function of \( \epsilon \) are reversed, i.e., the probability for filling limit orders increases as \( \epsilon \) increases. The crossover point where the fill probabilities are roughly the same occurs at \( p \approx p_c \). This is consistent with the depth profile in Figure 2.3 which also shows that depth profiles for different values of \( \epsilon \) cross at about \( p \approx p_c \).

Similarly, Figure 2.13 shows the average time \( \tau \) taken to fill an order placed at a distance \( p \) from the instantaneous mid-price. Again we see that though the average time is larger at larger values of \( \epsilon \) for small \( p/p_c \), this behaviour reverses at \( p \approx p_c \).

**2.2.3 Varying tick size \( dp/p_c \)**

The dependence on discrete tick size \( dp/p_c \) of the cumulative distribution function for the spread, instantaneous price impact, and mid-price diffusion, are shown in Figure 2.14. We chose an unrealistically large value of the tick size, with
Figure 2.12: The probability $\Gamma$ for filling a limit order placed at a price $p/p_c$ where $p$ is calculated from the instantaneous mid-price at the time of placement. The three cases correspond to Figure 2.3: $\epsilon = 0.2$ (solid), $\epsilon = 0.02$ (dash), $\epsilon = 0.002$ (dot).

Figure 2.13: The average time $\tau$ nondimensionalized by the rate $\delta$, to fill a limit order placed at a distance $p/p_c$ from the instantaneous mid-price.
$dp/p_c = 1$, to show that, even with very coarse ticks, the qualitative changes in behavior are typically relatively minor.

Figure 2.14(a) shows the cumulative density function of the spread, comparing $dp/p_c = 0$ and $dp/p_c = 1$. It is apparent from this figure that the spread distribution for coarse ticks “effectively integrates” the distribution in the limit $dp \to 0$. That is, at integer tick values the mean cumulative depth profiles roughly match, and in between integer tick values, for coarse ticks the probability is smaller. This happens for the obvious reason that coarse ticks quantize the possible values of the spread, and place a lower limit of one tick on the value the spread can take. The shift in the mean spread from this effect is not shown, but it is consistent with this result; there is a constant offset of roughly $1/2$ tick.

The alteration in the price impact is shown in Figure 2.14(b). Unlike the spread distribution, the average price impact varies continuously. Even though the tick size is quantized, we are averaging over many events and the probability of a price impact of each tick size is a continuous function of the order size. Large tick size consistently lowers the price impact. The price impact rises more slowly for small $p$, but is then similar except for a downward translation.

The effect of coarse ticks is less trivial for mid-price diffusion, as shown in Figure 2.14(c). At $\epsilon = 0.002$, coarse ticks remove most of the rapid short-term volatility of the midpoint, which in the continuous-price case arises from price fluctuations smaller than $dp/p_c = 1$. This lessens the negative autocorrelation of midpoint price returns, and reduces the anomalous diffusion. At $\epsilon = 0.2$, where both early volatility and late negative autocorrelation are smaller, coarse ticks have less effect. The net result is that the mid-price diffusion becomes less sensitive to the value of $\epsilon$ as tick size increases, and there is less anomalous price diffusion.
Figure 2.14: Dependence of market properties on tick size. Heavy lines are \( dp/p_c \to 0 \); light lines are \( dp/p_c = 1 \). Cases correspond to Figure 2.3, with \( \epsilon = 0.2 \) (solid), \( \epsilon = 0.02 \) (dash), \( \epsilon = 0.002 \) (dot). (a) is the cumulative distribution function for the nondimensionalized spread. (b) is instantaneous nondimensionalized price impact, (c) is diffusion of the nondimensionalized midpoint shift, corresponding to Figure 2.11.
Chapter 3

On the origin of large price changes

In Chapter 2 we have seen that a simplistic limit order based model is able to make predictions about many characteristics of the price formation. However, the model introduced previously fails to properly describe the price diffusion. It is unable to explain the weakness of the autocorrelation of the return time series. The model also underestimates the tail of the return distribution, even though it shows diffusion anomaly. Both of these properties contradict to the stylized facts. In this chapter we perform a real data analysis in order to gain insight into these problems.

3.1 Background

It has been known for more than forty years that price changes are fat-tailed [22, 23, 52, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63], i.e there is a higher probability of extreme events than in a normal distribution. This is a stylized fact of markets, and it has a strong implication for financial risk. There have been several conjectures about the origin of fat tails in prices [64, 66, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75]. Most of these theories are either generic mechanisms for generating power laws, such as multiplicative noise or maximization of alternative entropies or agent-based models that make qualitative predictions that are not very specific.

Two theories that deserve special mention because they make testable hypotheses about the detailed underlying mechanism are the subordinated random process theory of Clark [64] and the recent theory of Gabaix et al. [75]. Clark’s proposal is that because order arrival rates are highly intermittent, aggregating in a fixed time interval leads to fat tails in price returns. Gabaix et al.’s proposal is that high volume orders cause large price movements. We show that neither of these theories describes large price changes in the London Stock Exchange.

Instead, we show that large price fluctuations are driven by fluctuations in liquidity, i.e., variations in the response of prices to changes in supply and demand [76]. The number of agents that participate in the market at any given time, and
thus the number of orders to buy or sell, is rather small. Even for a heavily traded
stock, the typical number of orders on one side of the book at any given time is
generally around 30. While it is in some cases a good approximation to regard
the market as a statistical system, which can be treated using mathematical
methods from statistical mechanics [27, 50, 77, 29, 78, 79], markets are far from
the thermodynamic limit, and display strong finite size effects. Fluctuations
in orders are important, but it is not the size of orders that drives large price
changes, but rather the uniformity of their coverage of price levels. Revealed
supply and demand curves at any instant in time are irregular step-like functions
with long flat regions and large jumps. The market can be regarded as a granular
medium, in which the incremental changes in supply and demand are the grains.
The statistical properties of prices depend more on the fluctuations in revealed
supply and demand than on their mean behavior.

A common assumption is that large price returns \( r(t) \) are asymptotically
distributed as a power law. In more technical terms, letting \( m(t) \) be the midprice
at time \( t \), this means that \( r(t) = \log m(t) - \log m(t - \tau) \) satisfies \( P(r > x) \sim x^{-\alpha} \).
\( P(r > x) \) is the probability that \( r > x \), \( \tau \) is an arbitrary time interval, and
\( f(x) \sim g(x) \) means that \( f(x) \) and \( g(x) \) scale the same way\(^1\) in the limit \( x \to \infty \). \( \alpha \)
is called the tail exponent. It was initially thought that \( \alpha < 2 \), which is significant
because this would imply that the standard deviation of price returns does not
exist, and that under aggregation independent price returns should converge to
a Levy stable distribution [22, 52]. However, most subsequent studies indicate
that \( \alpha > 2 \) is more common [53, 54, 55, 56, 57, 58, 59, 61, 62, 63]. Nonetheless,
it still remains controversial whether a power law is always the best description
of price returns. Here, we do not attempt to resolve this debate. However, we
will use power laws as a useful way to describe the asymptotic behavior of price
changes, in particular to compare the distribution of price changes for different
stocks. Our results suggest that tail exponents vary from stock to stock, and are
not universal.

3.1.1 What causes returns?

Note that there is no unique notion of price in a real market. We need to dis-
tinguish the following notations. We let \( \pi \) be the limit price of a limit order,
and \( m(t) = (a(t) + b(t))/2 \) be the midpoint price or midprice defined by the
best quotes. All the results of this chapter concern the midprice, rather than
transaction prices, but at longer timescales this makes very little difference, since
the midpoint and transaction prices rarely differ by more than half the spread.
The midpoint price is more convenient to study because it avoids problems asso-
ciated with the tendency of transaction prices to bounce back and forth be-
tween the best bid and ask. Price changes are typically characterized as returns
\( r_\tau(t) = \log m(t) - \log m(t - \tau) \).

Here, we study changes in the midprice at the level of individual events. The

\(^1\)\( f(x) \sim g(x) \) means that \( \lim_{x \to \infty} L(x)f(x)/g(x) = 1 \), where \( L(x) \) is a slowly varying
function. A slowly varying function \( L(x) \) is a positive function that for every \( t \) satisfies
\( \lim_{x \to \infty} L(tx)/L(x) = 1 \).
arrival of three kinds of events can cause the midprice to change:

- **Market orders.** A market order bigger than the opposite best quote widens the spread by increasing the best ask if it is a buy order, or decreasing the best bid if it is a sell order.

- **Limit orders.** A limit order that falls inside the spread narrows it by increasing the best bid if it is a buy order, or decreasing the best ask if it is a sell order.

- **Cancellations.** A cancellation of the last limit order at the best price widens the spread by either increasing the best ask or decreasing the best bid.

We prefer to study individual events for several reasons: (1) It removes any ambiguity about time scale, and makes it easier to compare stocks with different activity levels. (2) It minimizes problems associated with clustered volatility (the positive autocorrelation of the absolute value of price changes). Clustered volatility is also driven by variations in event arrival rates, so its effect is weaker when the time lag between price changes is measured in terms of number of intervening events. (3) Individual events are the most fundamental level of description – returns on any longer scale can be built out of returns on an event scale. We will generally measure time in terms of event time: the time interval between events $t_1$ and $t_2$ is measured as the number of intervening events plus one. (One is added so that two adjacent events comprise an event time interval of one.)

Analysis at the event scale is particularly useful under the assumption that large price fluctuations are asymptotically power law distributed because of the invariance properties of power laws under aggregation. If two independent variables $X$ and $Y$ both have power law tails with exponents $\alpha_X$ and $\alpha_Y$, then the variable $Z = X + Y$ also has a power law tail with exponent $\min(\alpha_X, \alpha_Y)$. Thus, if we show that returns at the level of individual events have power law tails with exponent $\alpha$ and if there are not strong and persistent correlations, returns on any longer event time scale should also have power law tails with exponent $\alpha$. In fact, $\alpha$ appears to be the same whether returns are measured in event time or real time (see Section 4.4).

### 3.1.2 Review of previous work

There is considerable prior work using limit order data to address questions about market microstructure. For example, in a very early study Niederhoffer and Osborne discussed the granularity of revealed supply and demand, and showed orders tend to cluster at particular prices [80]. Clustering of limit orders, as well as stop-loss and take-profit orders are also studied in references [81, 82, 83]. Of particular relevance is the work of Biais, Hillion and Spatt [84] who, for the Paris Bourse, document several properties of order flow, including the concentration of orders near the best prices. They study the relation between order flow and the dynamics of prices, and mention the existence of gaps in the limit order book. Another relevant observation is made by Knez and Ready [85] and Petersen and Umlauf [86] who demonstrate that for the New York Stock
Exchange (NYSE) the most important conditioning variable for price impact is the size of an order relative to the volume at the best price. Goldstein and Kavajecz [87] document the effect of changes in tick size on limit order book depth, while several theoretical papers justify the existence of positive bid-ask spread and investigate the motivations for placing limit orders as opposed to market orders [88, 89, 90, 91, 92, 93]. In addition, we should mention recent empirical studies by Coppejans and Domowitz [94] and Coppejans, Domowitz and Madhaven [95] that document variations in liquidity and co-movements of liquidity with returns and volatility.

Our work adds to this literature by investigating the relationship between order placement and price movement at the level of individual events. Our motivation is to understand what drives large price movements, to gain insight into the fat tails of price returns. A relevant paper in this regard is the work of Plerou et al. [96], who showed that for the NYSE in a fixed time interval the scaling behavior in the standard deviation of individual price fluctuations roughly matches that of price fluctuations, and dominates over fluctuations in the number of events. This seems to contradict the later conclusions of the same authors in Gabaix et al. [75].

### 3.2 Data

In order to have a representative sample of high volume stocks we selected 16 companies traded on the London Stock Exchange (LSE) in the 4-year period 1999-2002. The stocks we analyzed are Astrazeneca (AZN), Baa (BAA), BHP Billiton (BLT), Boots Group (BOOT), British Sky Broadcasting Group (BSY), Diageo (DGE), Gus (GUS), Hilton Group (HG.), Lloyds Tsb Group (LLOY), Prudential (PRU), Pearson (PSON), Rio Tinto (RIO), Rentokil Initial (RTO), Reuters Group (RTR), Sainsbury (SBRY), Shell Transport & Trading Co. (SHEL). These stocks were selected because they have high volume and they are all continuously traded during the full period. Table 3.1 gives a summary of the number of different events for the 16 stocks\(^2\).

The London Stock Exchange consists of two parts: the completely automated electronic downstairs market (SETS) and the upstairs market (SEAQ). The trading volume is split roughly equally between the two markets. We study the downstairs market because we have a record of each action by each trader as it occurs. In contrast, trades in the upstairs market are arranged informally between agents, and are printed later. There are no designated market makers for SETS; however, any member of the exchange is free to act as a market maker by posting simultaneous bids and offers. This should be contrasted with the NYSE, for example, which has a designated specialist to trade each stock\(^3\). The book is

\(^2\)We have observed similar results on less liquid stocks, except that the statistics tend to be poorer, and as we discuss in Section 3.4, e.g., Figure 3.7, less liquid stocks appear to have lower tail exponents.

\(^3\)Another difference between the two markets is that clearing in the LSE is fully automated and instantaneous; in contrast, in the NYSE clearing is done manually, creating an uncertainty in response times.
Table 3.1: Summary statistics of the 16 stocks we studied for the period 1999-2002. The columns give the number of events of each type, in thousands. All events are “effective” events – see the discussion in Section 1.4.6.

Trading begins each day with an opening auction. There is a period leading up to the opening auction in which orders are placed, but no transactions take place. The market is then cleared and for the remainder of the day (except for occasional exceptional periods) there is a continuous auction. We remove all data associated with the opening auction, and analyze only orders placed during the continuous auction.

An analysis of the limit order placement shows that in our dataset approximately 35% of the effective limit orders are placed inside the book (\( \pi > a(t) \) or \( \pi < b(t) \)). 33% are placed at the best prices (\( \pi = b(t) \) or \( \pi = a(t) \)), and 32% are placed inside the spread (\( b(t) < \pi < a(t) \)). This is roughly true for all the stocks except for SHEL, for which the percentages are 71%, 18% and 11%, respectively. Moreover, for all the stocks the properties of buy and sell limit orders are approximately the same.

In this dataset cancellation occurs roughly 32% of the time at the best price.

---

<table>
<thead>
<tr>
<th>tick</th>
<th>market orders</th>
<th>limit orders</th>
<th>cancellations</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AZN</td>
<td>652</td>
<td>2,067</td>
<td>1,454</td>
<td>4,173</td>
</tr>
<tr>
<td>BAA</td>
<td>226</td>
<td>683</td>
<td>487</td>
<td>1,397</td>
</tr>
<tr>
<td>BLT</td>
<td>297</td>
<td>825</td>
<td>557</td>
<td>1,679</td>
</tr>
<tr>
<td>BOOT</td>
<td>246</td>
<td>711</td>
<td>501</td>
<td>1,458</td>
</tr>
<tr>
<td>BSY</td>
<td>404</td>
<td>1,120</td>
<td>726</td>
<td>2,250</td>
</tr>
<tr>
<td>DGE</td>
<td>527</td>
<td>1,329</td>
<td>854</td>
<td>2,709</td>
</tr>
<tr>
<td>GUS</td>
<td>244</td>
<td>734</td>
<td>518</td>
<td>1,496</td>
</tr>
<tr>
<td>HG.</td>
<td>228</td>
<td>676</td>
<td>472</td>
<td>1,377</td>
</tr>
<tr>
<td>LLOY</td>
<td>723</td>
<td>1,664</td>
<td>1,020</td>
<td>3,407</td>
</tr>
<tr>
<td>PRU</td>
<td>448</td>
<td>1,227</td>
<td>821</td>
<td>2,496</td>
</tr>
<tr>
<td>PSON</td>
<td>373</td>
<td>1,063</td>
<td>734</td>
<td>2,170</td>
</tr>
<tr>
<td>RIO</td>
<td>381</td>
<td>1,122</td>
<td>771</td>
<td>2,274</td>
</tr>
<tr>
<td>RTO</td>
<td>276</td>
<td>620</td>
<td>389</td>
<td>1,285</td>
</tr>
<tr>
<td>RTR</td>
<td>479</td>
<td>1,250</td>
<td>820</td>
<td>2,549</td>
</tr>
<tr>
<td>SBRY</td>
<td>284</td>
<td>805</td>
<td>561</td>
<td>1,650</td>
</tr>
<tr>
<td>SHEL</td>
<td>717</td>
<td>4,137</td>
<td>3,511</td>
<td>8,365</td>
</tr>
<tr>
<td>total</td>
<td>6,505</td>
<td>20,033</td>
<td>14,196</td>
<td>40,734</td>
</tr>
</tbody>
</table>

---

4In 2003 the LSE began to allow “iceberg orders”, which contain a hidden component that is only revealed as the exposed part of the order is removed.

5For the Paris Stock Exchange, from 1994 Biais, Hillion and Spatt [84] observed 42% of the orders inside the spread, 23% at the best, and 35% inside the book. We do not know why SHEL is anomalous, though it is worth noting that is the most heavily traded stock in the sample, a close proxy is also traded on the NYSE, and it has the second largest tail exponent among the 16 stocks chosen for the study.
and 68% of the time inside the book. This is quite consistent across stocks and between the cancellation of buy and sell limit orders. The only significant deviation is once again SHEL, for which the percentages are 14% and 86%.

### 3.3 Fluctuations in liquidity drive the tails

The assumption that large price changes are caused by large market orders is very natural. A very large market order will dig deeply into the limit order book, causing transactions at many price levels, increasing the spread, and changing the midprice. Surprisingly, this is *not* the cause of most large price changes. Instead, as we will demonstrate in this section, most large price changes are due to discrete fluctuations in liquidity, manifested by gaps in filled price levels in the limit order book. Large price changes caused by large orders are very rare, and play an insignificant role in determining the statistical properties of price changes.

In this section we will focus on price changes caused by market orders, and in Section 3.5 we will discuss price changes due to limit orders and cancellations.

#### 3.3.1 Large returns are not caused by large orders

We first demonstrate that most large returns are not caused by the arrival of large market orders. The probability density function of price returns that are caused by market orders can trivially be written as

\[ p(r) = \int p(r|\omega)p(\omega)d\omega, \]

where \( p(r) \), \( p(\omega) \) and \( p(r|\omega) \) are the probability density functions for returns \( r \), market order size \( \omega \), and returns given market order size. The conditional probability \( p(r|\omega) \) characterizes the response of prices to new orders, and can be viewed as the probability density of market impacts, or alternatively, as characterizing the distribution of liquidity for market orders. When a market order of size \( \omega \) arrives, the midprice will move if \( \omega \) is larger than or equal to the volume at the matching best price (i.e., the bid for sell market orders and the ask for buy market orders). In the limit of continuous prices we can trivially write

\[ p(r|\omega) = (1 - g(\omega))\delta(r) + g(\omega)f(r|\omega) \]

where \( \delta(.) \) is the Dirac delta function\(^6\) and \( g(\omega) \) is the probability that the midprice moves as a function of the order size \( \omega \). The function \( f(r|\omega) \) is the probability of a price shift \( r \), conditioned on the midprice moving in response to an order of volume \( \omega \). \( g(\omega) \) and \( f(r|\omega) \) behave quite differently. \( g(\omega) \) depends strongly on \( \omega \). For the LSE it scales roughly as \( g(\omega) \sim \omega^{0.3} \), about the same as the average market impact. In contrast, \( f(r|\omega) \) is surprisingly independent of the volume \( \omega \), in the sense that the unconditional fluctuations in \( r \) dominate the dependence on \( \omega \).

To demonstrate this in Figure 3.1 we show the cumulative probability for nonzero price returns conditioned on order size, i.e., \( F(r > X|\omega) = \int_X^\infty f(r|\omega)dr \) for several different ranges of market order size \( \omega \). The orders are sorted by size.

---

\(^6\)The Dirac delta function \( \delta(x) \) is defined so that \( \int \delta(x) = 1 \) over any domain that includes 0 and \( \int \delta(x) = 0 \) otherwise.
Figure 3.1: Dependence of returns on order size. $F(r > X | \omega)$ is the probability of a return $r > X$ conditioned on the order size $\omega$ and on the fact that the price shift is nonzero. Results shown are for buy orders, but similar results are seen for sell orders. The orders are sorted by size into five groups with roughly the same number of orders in each group. Ranging from small orders to large orders, the curves are black, red, green, blue, and magenta. In panel (a) we show the result for AZN and in panel (b) we show the average over the pool of 16 stocks described in Table 3.1. For the pooled data for each stock we normalize the order volume to the sample mean, and then combine the data. Each curve approximately approaches a power law for large $r$ independent of $\omega$, illustrating that the key property determining large price returns is fluctuations in market impact, and that the role of the volume of the order initiating a price change is minor.
Figure 3.2: $p(\omega|\omega > X)$, the distribution of market order sizes conditioned on generating a return greater than $X$, for the stock AZN. The values of $X$ correspond to the 50 percentile (red), 90 percentile (blue), and 99 percentile (green) of the return distribution. The black line is the unconditional distribution $p(\omega)$. The jumps are due to the tendency to place orders in round numbers of shares.

into five groups with roughly the same number of orders in each group. The distributions for each range of $\omega$ are roughly similar, both for individual stocks such as AZN, and for the pool. Each curve approximately approaches a power law for large $r$ independent of $\omega$, illustrating that the key property determining large price returns is fluctuations in market impact, and that the role of the volume of the order initiating a price change is minor. For the pooled stock result, for large returns the curves for large order size tend to be on top of those for small order size, illustrating a weak dependence on order size, but this is relatively small, and not visible in the results for individual stocks. Although we do not present them here, using the TAQ database we have obtained similar results for a small sample of stocks traded in the NYSE, suggesting that this behavior is not specific to the LSE\textsuperscript{7}. This is particularly interesting given the significant differences in the structure of the NYSE, and also because the TAQ data set includes upstairs trades.

To reinforce this point, in Figure 3.2 we show $p(\omega|\omega > X)$, the probability of a market order of size $\omega$ conditioned on the return being greater than a certain threshold $X$. We do this for the stock Astrazeneca (AZN) for several different values of $X$, getting virtually the same curve independent of $X$. The distribution of sizes of orders that generate large returns is essentially the same as the distri-

\textsuperscript{7}Based on five-minute averages, Weber and Rosenow [97] also see that liquidity is the dominant effect, using data from NYSE and Island.
bution of sizes of orders that generate small returns. Similar results are obtained for all the stocks in our sample. Thus, it seems that order size does not play an important role in generating large returns.

Gabaix et al. [75] have recently proposed that large returns can be explained by assuming the market impact density function $p(r|\omega)$ is sharply peaked around a central value $k\omega^{1/2}$, so that it can be approximated as a Dirac delta function $\delta(r - k\omega^{1/2})$. The result of Figure 3.1 makes it clear that the sharply peaked assumption is a poor approximation – the distribution is quite broad, in the sense that the conditional distribution $p(r|\omega) \approx p(r)$, and $p(r)$ is not sharply peaked.

### 3.3.2 Volume dependence of mean market impact

From a variety of previous studies it is clear that the mean market impact is an increasing function of $\omega$ [100, 33, 34, 35, 101, 102, 103, 104, 105, 98, 77]. How do we reconcile this with our claim that the distribution of market impact is almost independent of volume? The key is that the main effect of changing order volume is to change the probability that the price will move, with very little effect on how much it moves. From Eq. (1) one easily obtains

$$E(r|\omega) = g(\omega) \int f(r|\omega) r \, dr \propto g(\omega)$$

where for the last proportionality relationship we have used the result from Figure 3.1 that $f(r|\omega)$ is almost independent of $\omega$. Thus, the expected price change scales like the probability of a price change, a relationship that we have verified for both the LSE and the NYSE. However, this variation is still small in comparison with the intrinsic variation of returns; the mean market impact of a very large order is less than the average size of the spread, but the largest market impacts are often more than ten times this large.

### 3.3.3 Correlations between order size and liquidity

One possible explanation of the independence of price response and order size is that there is a strong correlation between order size and liquidity. There is an obvious strategic reason for this: agents who are trying to transact large amounts split their orders and execute them a little at a time, watching the book, and taking whatever liquidity is available as it enters. Thus, when there is a lot of volume in the book, they submit large orders, and when there is less volume, they submit small orders. This effect tends to even out the price response of large and small orders. We will see that this effect indeed exists, but it is only part of the story, and is not the primary determinant of the behavior we observe here.

---

8 Analyses of several markets make it clear that the mean response of prices to orders varies considerably from market to market, and is not in general well characterized by a square root law [98, 77, 78, 126]. It is also worth noting that the volume distribution in Figure 3.2 does not appear to have a power law tail. This is true of all the stocks in our sample. In contrast, power law tails are observed for stocks in the NYSE [99].
Table 3.2: Summary table of the percentage of the time that nonzero changes in the best prices are equal to the first gap (left) and that the market order volume $\omega$ exactly matches the volume at the corresponding best price $V_{\text{best}}$ (right). Assuming a Bernoulli process, the sample errors are the order of $0.1 - 0.2\%$.

<table>
<thead>
<tr>
<th>tick</th>
<th>% of nonzero return equal to first gap</th>
<th>% of nonzero return with $\omega = V_{\text{best}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sell</td>
<td>buy</td>
</tr>
<tr>
<td>AZN</td>
<td>94.3</td>
<td>99.6</td>
</tr>
<tr>
<td>BAA</td>
<td>95.9</td>
<td>99.1</td>
</tr>
<tr>
<td>BLT</td>
<td>95.0</td>
<td>99.2</td>
</tr>
<tr>
<td>BOOT</td>
<td>95.9</td>
<td>99.2</td>
</tr>
<tr>
<td>BSY</td>
<td>94.3</td>
<td>99.7</td>
</tr>
<tr>
<td>DGE</td>
<td>96.3</td>
<td>99.7</td>
</tr>
<tr>
<td>GUS</td>
<td>95.8</td>
<td>99.5</td>
</tr>
<tr>
<td>HG.</td>
<td>95.9</td>
<td>99.5</td>
</tr>
<tr>
<td>LLOY</td>
<td>97.3</td>
<td>99.8</td>
</tr>
<tr>
<td>PRU</td>
<td>95.9</td>
<td>99.5</td>
</tr>
<tr>
<td>PSN</td>
<td>93.1</td>
<td>99.6</td>
</tr>
<tr>
<td>RIO</td>
<td>95.7</td>
<td>99.7</td>
</tr>
<tr>
<td>RTO</td>
<td>96.1</td>
<td>99.5</td>
</tr>
<tr>
<td>RTR</td>
<td>93.0</td>
<td>99.7</td>
</tr>
<tr>
<td>SBRY</td>
<td>95.6</td>
<td>99.4</td>
</tr>
<tr>
<td>SHEL</td>
<td>98.7</td>
<td>99.9</td>
</tr>
<tr>
<td>average</td>
<td>95.6</td>
<td>99.5</td>
</tr>
</tbody>
</table>

In fact, the unconditional correlation between market order size and volume at the best is rather small. For AZN, for example, it is about 1%. However, if we restrict the sample to orders that change the midprice, the correlation soars to 86%. The reason for this is that for orders that do not change the price, there is essentially no correlation between order size and volume at the best. For the rarer case of orders that do change the price, in contrast, most market orders exactly match the volume at the best. As shown in Table 3.2, for the stocks in our sample, 86% of the buy orders that change the price exactly match the volume at the best price. (This is 85% for sell orders.)

The relationship between the volume of market orders and the best price becomes more evident with a nonlinear analysis. Figure 3.3 shows $E(\omega|V_{\text{best}})$, where $\omega$ is the market order size and $V_{\text{best}}$ is the volume at the corresponding best price. We see in this figure that the expected order size is nonzero even for the smallest values of $V_{\text{best}}$. It grows monotonically with $V_{\text{best}}$, but with a slope that is substantially less than one, and a roughly concave shape. This makes the nonlinear correlation between order size and liquidity clear. However, in the following section we will see that the dependence of order size on liquidity is not strong enough to substantially suppress large price fluctuations.
Figure 3.3: The dependence of order size on liquidity. Liquidity is measured as the volume at the best price, which is binned into deciles with roughly equal numbers of events. The vertical axis shows the mean market order volume for each decile, and the horizontal shows the volume at the best. The units of both axes are thousands of shares. The ranges shown around each point indicate one standard deviation (and are not standard errors). The dashed line has slope 1 and serves as a point of comparison.
3.3.4 Liquidity fluctuations drive price fluctuations

In this section we demonstrate in concrete terms that price fluctuations are driven almost entirely by liquidity fluctuations. To do this we study the virtual market impact, which is a useful tool for probing the supply and demand curves defined by the limit order book. Whereas the true market impact \( p(\tau|\omega) \) tells us about the distribution of impacts of actual market orders, as discussed in Section 3.3.1, the virtual market impact is the price change that would occur at any given time if a market order of a given size were to be submitted. More formally, at any given time \( t \) the limit orders stored in the order book define a revealed supply function \( S(\pi, t) \) and revealed demand function \( D(\pi, t) \). Let \( V(\pi, t) \) be the total volume of orders stored at price \( \pi \). The revealed supply function is

\[
S(\pi, t) = \sum_{i=a(t)}^{i=\pi} V(i, t)
\]  

The revealed supply function is non-decreasing, and so for any fixed \( t \) has a well defined inverse \( \pi(S, t) \). The virtual market impact is the price shift caused by a hypothetical order of size \( S \), e.g., for buy orders it is \( \pi(S, t) - a(t) \). The virtual market impact for sell market orders can be defined in terms of the revealed demand in a similar manner. By sampling at different values of \( t \), for any fixed hypothetical order size we can create a sample distribution of virtual market impacts. This naturally depends on the sampling times, but these can be chosen to match any given set of price returns.

In Figure 3.4 we show the cumulative distribution of virtual market impacts for the stock AZN for several different values of \( D \), corresponding to different quantiles of market order size. The cumulative distributions define a set of approximately parallel curves. They are shifted as one would expect from the fact that larger hypothetical order sizes tend to have larger virtual market impacts. These curves are similar in shape to the distribution of returns. Most striking, for the median market order size, the curves are almost identical. We see similar results for all the stocks in our sample. This demonstrates quite explicitly that the distribution of returns is determined by properties of the limit order book, and that the typical price return corresponds to the price response to an order of typical size. The fact that the price distribution is so close to the virtual market impact of a typical order shows that correlations between order size and liquidity are not important in determining price fluctuations.

3.4 Granularity of supply and demand

At first sight, the behavior described in the previous section seems baffling: how can market order size be so unimportant to price response? In this section we show how this is due to granularity of revealed supply and demand, which causes large fluctuations in liquidity.

The cause of this puzzling behavior is fluctuations in occupied price levels in the limit order book. In particular, one can define the size of the first gap \( g \) as the
Figure 3.4: The cumulative probability distribution $P(r > x)$ of virtual market impacts for the stock AZN, based on samples of the limit order book taken just before the arrival of a market order. From left to right the continuous lines are the virtual market impact for hypothetical buy orders of size $\omega = 100$ shares (red), $\omega = 1,600$ shares (green), $\omega = 8,700$ shares (blue), and $\omega = 25,000$ shares (magenta); these correspond to the 0.1, 0.5, 0.9, and 0.99 quantiles of market order size. These are compared to the average distribution of returns (black dashed), which is very similar to the virtual market impact for the 0.5 quantile. Note that we have discarded cases in which the virtual market impact is undefined due to excessively large hypothetical market order volume, but these comprise only about 3% of the events.
Figure 3.5: A typical configuration of the limit order book for AZN before and after a large price fluctuation. The two panels plot the volume (in shares) of limit orders at each price level; sell limit orders are shown as positive, and buy limit orders as negative. In panel (a) we see that there is a large gap between the best ask price and the next highest occupied price. The arrival of a market order to buy removes all the volume at the best ask, giving the new limit order book configuration shown in panel (b), which has a much higher best ask price than previously.

The absolute difference between the best log price $\pi_{\text{best}}$ and the log price of the next best quote, $\pi_{\text{next}}$ as $g = |\log \pi_{\text{best}} - \log \pi_{\text{next}}|$. In Figure 3.5 we show a typical set of events before and after a large price fluctuation. In panel (a) we see that there is a large first gap on the sell side of the limit order book. In panel (b) we see the configuration of the book an event later, after a market order of exactly the same size has removed all the volume at the best ask. This results in a large change in the midpoint price. Thus, the large return simply reflects a large gap inside the book that has been revealed by the removal of the best ask. We find that this is the typical behavior underlying almost all large returns.

As already shown in Table 3.2, typically about 85% of market orders that result in price changes exactly match the volume at the best price. Furthermore, as shown in the first two columns of this table, when they do exceed the volume at the best price, it is quite rare that they penetrate the next occupied level. For buy (sell) market orders, on average 95.5% (99.5%) of the nonzero shifts in the best price are exactly equal to the first gap. Interestingly there is a significant asymmetry between buy and sell. Traders act to minimize their transaction costs, so that jumps of more than one gap are rare, particularly for sell orders. (However, as we stressed earlier, the return distribution can be generated by a constant order of median size – this correlation is interesting, but not essential.) Note that the trader initiating the change does not pay a large spread – that would only happen to the next trader, if she were to immediately place a market order$^9$.

$^9$In general after a large shift in the bid or ask price, the next orders tend to be limit orders, but we have not yet been able to study the statistical properties of the sequence of subsequent events in detail. See [88, 84, 92, 93], and [86] for an empirical study of the role of quote size
This is by far the most common scenario that generates large price changes. In Figure 3.6(a) we compare the distribution of the first gaps to the distribution of price returns for the stock AZN. We see that the distributions are very similar. For the first gap size the tail index $\alpha = 2.52 \pm 0.07$, and for the return distribution $\alpha = 2.57 \pm 0.08$, showing that the scaling behaviors are similar. However, the similarity is not just evident in the scaling behavior – the match is good throughout the entire range, illustrating that most of the large price changes are caused by events of this type. Panel (b) of Figure 3.6 shows the same comparison for the pool of 16 stocks. The agreement in this case is even more striking\textsuperscript{10}.

To demonstrate that the correspondence in the above figure is not just a coincidence for AZN, we have computed the tail exponents for returns and first gap size for all 16 stocks in our data set. We do this using a Hill estimator\textsuperscript{[106]} by considering the largest $\sqrt{n}$ returns, where $n$ is the size of the sample. The results are shown in Figure 3.7, where we plot the tail exponents of returns against those of first gaps. The points cluster tightly along the diagonal, making it clear that there is a strong positive correlation ($R^2 = 0.93$) between the gap tail exponents and the return tail exponents; a least squares linear fit has a slope of $0.98 \pm 0.05$, and we are unable to reject the null hypothesis that the tail exponents of the gaps and returns are drawn from the same distribution.

It is worth noting that lightly traded stocks tend to have a smaller tail exponent than heavily traded stocks. This is already clear in Figure 3.7, where we have used a color coding to identify stocks of different volumes. To investigate this more quantitatively, in Figure 3.8 we plot the tail exponent of price changes against the number of market orders for each stock in the sample. While the relationship is noisy, there is a clear positive trend; the slopes of the linear fits to positive and negative returns are both highly statistically significant. The tail exponents vary from about 1.6 to 2.8, whereas the error bars are more than a factor of ten smaller than the range of variation\textsuperscript{11}. This suggests that the tail exponent is not universal. Not surprisingly, stocks that are more heavily traded tend to display less extreme risk than stocks that are lightly traded.

### 3.5 Limit orders and cancellations

We have so far considered only returns caused by market orders. In this section we will discuss returns caused by limit orders and cancellations, and show that they are statistically indistinguishable from those caused by market orders.

The distribution of returns caused by limit orders, market orders, and cancellations for AZN is shown in Figure 3.9. The distributions are quite similar.

---

\textsuperscript{10}We have also studied the distribution of higher order gaps. Moving away from the best price $\pi_0$, the $n^{th}$ order gap for $n = 1, 2, \ldots$ can be recursively defined as $g_n = |\log \pi_{n-1} - \log \pi_n|$, where $\pi_n$ is the $n^{th}$ occupied price level. Interestingly, we find that the tail behavior of higher order gaps is the same as that of $g_1$.

\textsuperscript{11}Table 3.3 gives error bars for tail exponent estimations based on absolute returns (including both buy and sell events), while Figures 3.6 and 3.7 are based on buy and sell returns taken separately, but the error bars are comparable.
Figure 3.6: The cumulative distribution $P(g > x)$ of the size of first gaps $g$ (red continuous line), compared to the cumulative distribution of returns generated by market orders $P(r > x)$ (black dashed line). Panel (a) refers to buy market orders for AZN, in double logarithmic scale to highlight the tail behavior. The two distributions are very similar. The result is even more impressive when we consider the average over the 16 stocks described in Table 3.1 (Panel (b)).
Figure 3.7: A comparison of tail exponents $\alpha$ for returns (horizontal axis) vs. tail exponents for the first gap (vertical axis). The first gaps are sampled immediately preceding the market order that triggers the return. This is done for all the stocks in Table 3.1, and includes both the buy and sell side of the limit order book. We use red circles for low liquidity stocks, blue squares for medium liquidity stocks, and green triangles for high liquidity stocks. Empty symbols refer to sell market orders and filled symbols to buy market orders. The black dashed line is the linear regression. We see that there is a clear positive relationship between the number of orders and the tail exponent.
Figure 3.8: Dependence of the tail exponent of price changes on the number of market orders, for the data of Figure 3.7. Black circles are for positive returns (caused by buy market orders), and red triangles are for negative returns (caused by sell market orders). The steeper red curve is the fit to negative returns and the black curve the fit to positive returns.

Figure 3.9: A comparison of the distribution of absolute returns (including both buy and sell orders) caused by limit orders, market orders, and cancellations for AZN. The black circles are for market orders, the blue triangles are for limit orders, and the red squares are for cancellations.
Table 3.3: Hill estimator of the return caused by market orders, limit orders and cancellations. We analyzed the top percentile and give 95% confidence intervals. The confidence intervals assume the samples are uncorrelated, and so may be over-optimistic.

To make the analysis more quantitative, we compute the tail exponents for the 16 stocks in the sample. Table 3.3 shows the Hill estimates for returns caused by market orders, limit orders, and cancellations. The table shows clearly that the tail exponents for the three distributions are very close: for all 16 stocks the estimated tail exponents are well within the 95% confidence intervals. On the other hand, Table 3.3 shows that there is a much larger variation across stocks.

Why are the distributions of different events so similar? The correspondence between returns caused by market orders and returns caused by cancellations is not surprising. In terms of the effect on best prices, removal of the volume at the best price by cancellation is equivalent to removal by a market order, in both cases creating a price change equal in size to the first gap. However, for a limit order this is not so obvious: a limit order that falls inside the spread decreases the spread, and causes a price change in the opposite direction from market orders and cancellations. To investigate this we have studied the distribution of the spread, which also appears to be a power law, but with a larger tail exponent. At this point the reason for the close correspondence between the returns generated by limit and market orders remains unexplained.

### 3.6 Correlations in occupied sites

A natural hypothesis about the origin of the large gap distributions involves fluctuations in the number of occupied sites in the order book. If there are only
a few orders in the book, so that it is mostly empty, we naturally expect large
gaps to exist. Perhaps the approximate power for the gap distribution can be
explained by a similar power law in the number of occupied sites? In this section
we show that this is not the case, but rather that the power law behavior depends
on correlations in occupied price levels.

Interestingly, we do find evidence of power law scaling in the number of oc-
cupied price levels. Figure 3.10 (a) shows the probability density function \( p(N) \),
where \( N \) is the number of occupied price levels in the order book at any given
time. The low \( N \) region is well fit by a power-law \( p(N) \sim N^\beta \) over about two
decades.

However, the fluctuations in the number of orders stored in the book cannot
explain the power law behavior of prices, because the tail exponent of the power
law behavior of the gap distribution near zero is much too large. If we assume
that the location of order deposition is uncorrelated, then for order arrival on a
bounded domain gap size should be proportional to \( 1/N \). But with \( p(N) \sim N^\beta \)
and \( g \sim 1/N \), \( p(g) \sim g^{-(\beta+2)} \). For \( \beta \approx 5.5 \) this results in \( p(g) \sim g^{-7.5} \). In
contrast, the empirically observed scaling exponent for the density for AZN is
about 3.5.

Surprisingly, the distribution of gap size shows approximate power law scaling,
independent of \( N \). This is illustrated for AZN in Figure 3.10(b), where we plot
the probability of gap size conditioned on \( N \), \( p(g|N) \), for several different values
of \( N \). The same approximate power law scaling behavior is seen independent of
\( N \). The approximate power law behavior is evident even with as many as 60
occupied sites on one side of the book. Considering that it is rare for orders to
be placed more than 100 price ticks away from the best price, this illustrates that
occupied sites display nontrivial correlations, which are essential for explaining
large price fluctuations. This conclusion is reinforced by studies we have done
comparing the real distribution of occupied levels with models based on IID order
placement, which do not reproduce the power law behavior.
Figure 3.10: The effect of the finite number of orders for the stock AZN. Panel (a) shows the unconditional distribution of the number of occupied price levels for buy orders, plotted on a double logarithmic scale. The dashed line is the best fit of the low $N$ region with a functional form $p(N) \propto N^\beta$. Panel (b) shows the probability density of first gap sizes conditional on the number of occupied price levels, $p(g|N)$. We divide the sample in four subsamples according to the value of $N$. Specifically, we have $0 < N \leq 15$ (black circles), $15 < N \leq 30$ (red squares), $30 < N \leq 45$ (green diamonds), and $45 < N \leq 60$ (blue triangles). Surprisingly, the first gap size shows approximate power law behavior even when the number of occupied sites is very large.
Chapter 4

Stylized facts on longer time scales: where do they come from?

4.1 Background

Besides the heavy tails of return distributions, the clustered volatility in price fluctuations is an important problem in financial economics\(^1\). Heavy tails means that large price fluctuations are much more common than they would be for a normal distribution, and clustered volatility means that the size of price fluctuations has strong positive autocorrelations, so that there are periods of large or small price change. Understanding these phenomena has considerable practical importance for risk control and option pricing. Although the cause is still debated, the view has become increasingly widespread that in an immediate sense both of these features of prices can be explained by fluctuations in volume, particularly as reflected by the number of transactions. In this chapter we show that while fluctuations in volume or number of transactions do indeed affect prices, they do not play the dominant role in determining either clustered volatility or heavy tails [108].

The model that is now widely believed to explain the statistical properties of prices has its origins in a proposal of Mandelbrot and Taylor [114] that was developed by Clark [64]. 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 Brownian motion, and \(\tau(t)\) is a stochastic time clock whose increments are IID and uncorrelated with \(X\). Clark hypothesized that the time clock \(\tau(t)\) is the cumulative trading volume in time \(t\). Since then several empirical studies have demonstrated a strong correlation be-

\(^{1}\)The observation that price fluctuations display heavy tails and clustered volatility is quite old. In his 1963 paper Mandelbrot notes that prior references to observations of heavy tails in prices date back to 1915. He also briefly discusses clustered volatility, which he says was originally pointed out to him by Hendrik Houthakker. The modern upsurge of interest in clustered volatility stems from the work of Engle [107].
tween volatility – measured as absolute or squared price changes – and volume [109, 110, 111, 112, 113]. More recently, evidence has accumulated that the number of transactions is more important than their size [115, 116, 117]. We show that volatility is still very strong even if price movements are recorded at intervals containing an equal number of transactions, and that the volatility observed in this way is highly correlated with volatility measured in real time. In contrast, if we shuffle the order of events, but control for the number of transactions so that it matches the number of transactions in real time, we observe a much smaller correlation to real time volatility. We interpret this to mean that the number of transactions is less important than other factors.

Several studies have shown that the distribution of price fluctuations can be transformed to a normal distribution under a suitable rescaling transformation. There are important differences between the ways this is done in these papers. Ane and Geman, [117] claim that the price distribution can be made normal based on a transformation that depends only on the transaction frequency. The other studies, in contrast, use normalizations that also depend on the price movements themselves. Plerou et al. [96] divide the price movement by the product of the square root of the transaction frequency and a measure of individual price movements, and Andersen et al. [118] divide by the standard deviation of price movements. We find that, contrary to Ane and Geman, it is not sufficient to normalize by the transaction frequency, and that in fact in many cases this hardly alters the distribution of returns at all. A related theory is due to Gabaix et al. [75], who have proposed that the distribution of the size of large price changes can be explained based on the distribution of volume and the volume dependence of the price impact of trades. We examine this hypothesis and show that when prices are sampled to hold the volume constant, the distribution of price changes, and in particular its tail behavior, is similar to that in real time. We present some new tests of their theory that give insight into why it fails. Finally, we study the long-memory properties of volatility and show that neither volume nor number of transactions can be the principal causes of long-memory behavior.

Note that when we say a factor “causes” volatility, we are talking about proximate cause as opposed to ultimate cause. For example, changes in transaction frequency may cause changes in volatility, but for the purposes of this study we do not worry about what causes changes in transaction frequency. It may be the case that causality also flows in the other direction, e.g., that volatility also causes transaction frequency. All we test here are correlations. In saying that transaction frequency or trading volume cannot be a dominant cause of volatility, we are making the assumption that if the two are not strongly correlated there cannot be a strong causal connection in either direction. We recognize that if the relationship between two variables is highly nonlinear it is possible to have a strong causal connection even if the variables are uncorrelated. Thus, in claiming a lack of causality, we are assuming that nonlinearities in the relationship between the variables are small enough that correlation is sufficient.

This chapter is organized as follows: in Section 4.2 we describe the data sets and the market structure of the LSE and NYSE. In Section 4.3 we define volume and transaction time and discuss how volatility can be measured in each
of them, and discuss the construction of surrogate data series in which the data are scrambled, but either volume or number of transactions is held constant. We then use both of these tools to study the relationship between volume and number of transactions and volatility. In Section 4.4 we discuss the distributional properties of returns and explain why our conclusions differ from those of Ane and Geman and Gabaix et al. In Section 4.5 we study the long-memory properties of volatility, and show that while fluctuations in volume or number of transactions are capable of producing long-memory, other factors are more important.

4.2 Data

This part of the study is based on data from two markets, the New York Stock Exchange (NYSE) and the London Stock Exchange (LSE). For the NYSE we studied two different subperiods, the 626 trading day period from Jan 1, 1995 to Jun 23, 1997, labeled NYSE1, and the 734 trading day period from Jan 29, 2001 to December 31, 2003, labeled NYSE2. These two periods were chosen to highlight possible effects of tick size. During the first period the average price of the stocks in our sample was $73.8 and the tick size was an eighth of a dollar, whereas during the second period the average price was $48.2 and the tick size was a penny. The tick size in the second period is thus substantially smaller than that in the first, both in absolute and relative terms. For each data set we have chosen 20 stocks with high capitalizations. The tickers of the 20 stocks in the NYSE1 set are AHP, AIG, BMY, CHV, DD, GE, GTE, HWP, IBM, JNJ, KO, MO, MOB, MRK, PEP, PFE, PG, T, WMT, and XON, and the tickers of the stocks in the NYSE2 set are AIG, BA, BMY, DD, DIS, G, GE, IBM, JNJ, KO, LLY, MO, MRK, MWD, PEP, PFE, PG, T, WMT, and XOM. There are a total of about 7 million transactions in the NYSE1 data set and 36 million in the NYSE2 data set.

For the LSE, we use a data set that differs from the one we studied in Chapter 3. We studied the period from May 2000 to December 2002, which contains 675 trading days. we made this choice because there was no overall change in tick size during this period for LSE stocks. The average price of a stock in the sample is about 500 pence, though the price varies considerably – it is as low as 50 pence and as high as 2500 pence. The tick size depends on the stock and the time period, ranging from a fourth of a pence to a pence. As for the NYSE, we chose stocks that are heavily traded. The tickers of the 20 stocks in the LSE sample are AZN, REED, HSBA, LLOY, ULVR, RTR, PRU, BSY, RIO, ANL, PSON, TSCO, AVZ, BLT, SBRY, CNA, RB., BASS, LGEN, and ISYS. In aggregate there is a total of 5.7 million transactions during this period, ranging from 497 thousand for HSBA to 181 thousand for RB.

For the NYSE we use the Trades and Quotes (TAQ) data set. This contains the prices, times, and volume of all transactions, as well as the best quoted prices to buy or sell and the number of shares available at the best quotes at the times when these quotes change. Transaction prices from the upstairs and downstairs markets are mixed together and it is not possible to separate them. Our analysis
in the NYSE is based on transaction prices, though we have also experimented with using the average price of the best quotes. For the 15 minute time scale and the heavily traded stocks that we study here, this does not seem to make very much difference.

The data set for the LSE contains every order and cancellation for the on-book exchange, and a record of all transactions in both the on-book and off-book exchanges. We are able to separate the on-book and off-book transactions. The on-book transactions have the advantage that the timing of each trade is known to the nearest second, whereas, as already mentioned, the off-book transactions may have reporting delays. Also, the on-book transactions are recorded by a computer, whereas the off-book transactions depend on human records and are more error prone. Because of these problems, unless otherwise noted, we use only the downstairs prices. Our analysis for the LSE is based entirely on mid-quote prices in the on-book market, defined as the average of the best quote to buy and the best quote to sell.

To simplify our analysis, we paste the time series for each day together, omitting any price changes that occur outside of the period of our analysis, i.e., when the market is closed. For the NYSE we omit the first few and last few trades, but otherwise analyze the entire trading day, whereas for the LSE we omit the first and last half hour of each trading day. We don’t find that this makes very much difference – except as noted in Section 4.4.3, our results are essentially unchanged if we include the entire trading day.

4.3 Volatility under alternative time clocks

4.3.1 Alternative time clocks

A stochastic time clock $\tau(t)$ can speed up or slow down a random process and thereby alter the distribution or autocorrelations of its discrete increments. When the increments of a stochastic time clock are IID it is called a subordinator, and the process $Y(t) = X(\tau(t))$ is called a subordinated random process. The use of stochastic time clocks in economics originated in the business literature [120], and was suggested in finance by Mandelbrot and Taylor [114] and developed by Clark [64]. Transaction time $\tau_\theta$ is defined as

$$\tau_\theta(t_i) = \tau_\theta(t_{i-1}) + 1,$$

where $t_i$ is the time when transaction $i$ occurs. Another alternative is the volume time $\tau_v$, defined as

$$\tau_v(t_i) = \tau_v(t_{i-1}) + V_i,$$

where $V_i$ is the volume of transaction $i$.

The use of transaction time as opposed to volume time is motivated by the observation that the average price change caused by a transaction, which is also called the average market impact, increases slowly with transaction size. This suggests that the number of transactions is more important than their size [100, 115, 116, 98].
Note that alternative time clocks measures significantly changes the lengths of the analyzed time periods for the different stocks. Higher capitalized stocks tend to be more liquid, that is, the total number of transactions and the total traded volume increases with company size [122, 121].

4.3.2 Volatility in transaction time

The main results of this study were motivated by the observation that clustered volatility remains strong even when the price is observed in transaction time. Letting $t_i$ denote the time of the $i^{th}$ transaction, we define the transaction time volatility\(^3\) for interval $(t_{i-K}, t_i)$ as $\nu_\theta(t_i, K) = |p(t_i) - p(t_{i-K})|$, where $p(t_i)$ is the logarithm of the midquote price at time $t_i$. Here and in all the other definitions of volatility we always use non-overlapping intervals, $p$ will refer to the logarithm of the price, and when we say “return” we always mean log returns, i.e., $r(t) = p(t) - p(t - \Delta t)$. A transaction time volatility series for the LSE stock Astrazeneca with $K = 24$ transactions is shown in Figure 4.1. For $K = 24$ the average duration is 15 minutes, though there are intervals as short as 14 seconds and as long as 159 minutes\(^4\). In Figure 4.1 there are obvious periods of high and low volatility, making it clear that volatility is strongly clustered, even through we are holding the number of transactions in each interval constant.

For comparison in Figure 4.2(a) we plot the real time volatility $\nu(t, \Delta t) = |p(t) - p(t - \Delta t)|$ based on $\Delta t = $ fifteen minutes, over roughly the same period shown in Figure 4.1. In Figure 4.2(b) we synchronize the transaction time volatility with real time by plotting it against real time rather than transaction

\(^3\)We use the absolute value of end-to-end price changes rather than the standard deviation to avoid sampling problems when comparing different time clocks. We use the midquote price for the LSE to avoid problems with mean reversion of prices due to bid-ask bounce. We find that on the time scales we study here this makes very little difference. For convenience in the NYSE data series we use the transaction price.

\(^4\)Remember that we are omitting all price changes outside of trading hours.
time. It is clear from comparing panel (a) and panel (b) that there is a strong contemporary correlation between real time and transaction time volatility.

As a further point of comparison in panel (c) we randomly shuffle transactions. We do this so that we match the number of transactions in each real time interval, while preserving the unconditional distribution of returns but destroying any temporal correlations. Let the (log) return corresponding to transaction \( i \) be defined as \( r(t_i) = p(t_i) - p(t_{i-1}) \). The shuffled transaction price series \( \tilde{p}(t_i) \) is created by setting \( \tilde{r}(t_i) = r(t_j) \), where \( r(t_j) \) is drawn randomly from the entire time series without replacement. We then aggregate the individual returns to create a surrogate price series \( \tilde{p}(t_i) = \sum_{k=1}^{K} \tilde{r}(t_k) + p(t_0) \). We define the shuffled transaction real time volatility as \( \tilde{\nu}(t, \Delta t) = |\tilde{p}(t) - \tilde{p}(t - \Delta t)| \). The name emphasizes that even though transactions are shuffled to create a surrogate price series, the samples are taken at uniform intervals of real time and the number of transactions matches that in real time. The shuffled transaction real time volatility series shown in Figure 4.2(c) is sampled at fifteen-minute intervals, just as for real time. The resulting series still shows some clustered volatility, but the variations are smaller and the periods of large and small volatility do not match the real time volatility as well. This indicates that the number of transactions is less important in determining volatility than other factors, which persist even when the number of transactions is held constant.

4.3.3 Correlations between transaction time and real time

To measure the influence of the number of transactions on prices more quantitatively we compare the correlations between the volatilities of these series. To deal with the problem that the sampling intervals in real time and transaction time are different we first convert both the real time and transaction time volatility series, \( \nu(t_i, K) \) and \( \nu_\theta(t_i, K) \) to continuous time using linear interpolation within each interval \( (t_{i-K}, t_i) \). This allows us to compute \( \rho(\nu, \nu_\theta) \), the correlation between real time and transaction time volatilities as if they were continuous functions. For comparison we also compute \( \rho(\nu, \tilde{\nu}_\theta) \), the correlation between the real time and the shuffled transaction real time volatility series. This is done for each stock in all three data sets. The results are plotted in Figure 4.3 and summarized in Table 4.1. In every case \( \rho(\nu, \tilde{\nu}_\theta) \) is significantly smaller than \( \rho(\nu, \nu_\theta) \). The correlations between real time volatility and transaction time volatility are between 35% and 49%, whereas the correlation with shuffled transaction real time volatility is between 6% and 17%. The ratio is typically about a factor of four. This demonstrates that fluctuations in transaction rates account for only a small portion of the variation in volatility. This quantitatively confirms what the eye sees in Figure 4.2 – the principal influences on volatility are still present even if the number of transactions is held constant. The way in which prices respond to transactions is much more important than the number of transactions.

4.3.4 Correlations between volume time and real time

We can test the importance of volume in relation to volatility in a similar manner. To do this we create a volatility series \( \nu_v \) sampled in volume time. Creating such
Figure 4.2: A comparison of three different volatility series for the LSE stock Astrazeneca. In panel (a) the volatility is measured over real time intervals of fifteen minutes. In panel (b) the transaction time volatility is measured at intervals of 24 transactions, which roughly correspond to fifteen minutes, and then plotted as a function of real time. Panel (c) shows the volatility of a price series constructed by preserving the number of transactions in each 15 minute interval, but drawing single transaction returns randomly from the whole time series and using them to construct a surrogate price series. The fact that the volatilities in (a) and (b) agree more closely than those in (a) and (c) suggests that fluctuations in transaction frequency plays a smaller role in determining volatility than other factors. Time is measured in hours and the period is the same as in Figure 4.1.
Figure 4.3: An illustration of the relative influence of number of transactions and volume on volatility. The correlation between real time volatility $\nu$ and various alternative volatilities $\nu_x$ is plotted against the correlation with the corresponding volatility $\tilde{\nu}_x$ when the order of events is shuffled. Each mark corresponds to a particular stock, data set, and time clock. The solid marks are for the transaction related quantities, i.e., $\nu_x = \nu_\theta$ and $\tilde{\nu}_x = \tilde{\nu}_\theta$, and the open marks are for volume related quantities, i.e., $\nu_x = \nu_v$ and $\tilde{\nu}_x = \tilde{\nu}_v$. Thus, the horizontal axis represents the correlation with volume or transaction time, and the vertical axis represents the correlation with shuffled volume or shuffled transaction real time. Stocks in the LSE are labeled with black circles, in the NYSE1 (1/8 tick size) with red squares, and the NYSE2 (penny tick size) with blue diamonds. Almost all the data lie well below the identity line, demonstrating that volume or the number of transactions has a relatively small effect on volatility – most of the information is present when these are held constant.
data set  \( E[\rho(\nu, \nu_0)] \)  \( E[\rho(\nu, \tilde{\nu}_0)] \)  \( E[\rho(\nu, \nu_v)] \)  \( E[\rho(\nu, \tilde{\nu}_v)] \)  \( \frac{E[\rho(\nu, \nu_0)]}{E[\rho(\nu, \nu_v)]} \)  \( \frac{E[\rho(\nu, \tilde{\nu}_0)]}{E[\rho(\nu, \tilde{\nu}_v)]} \)  

<table>
<thead>
<tr>
<th></th>
<th>LSE</th>
<th>NYSE1</th>
<th>NYSE2</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSE</td>
<td>49 ± 3</td>
<td>12 ± 4</td>
<td>43 ± 3</td>
</tr>
<tr>
<td>NYSE1</td>
<td>45 ± 6</td>
<td>17 ± 3</td>
<td>38 ± 6</td>
</tr>
<tr>
<td>NYSE2</td>
<td>35 ± 4</td>
<td>6 ± 2</td>
<td>26 ± 3</td>
</tr>
</tbody>
</table>

Table 4.1: Average correlations and ratios of average correlations between real time and alternate volatilities. Correlations are given as percentages, and error bars are standard deviations for the data shown in Figure 4.3. All the correlations are against the real time volatility \( \nu \). \( E[x] \) denotes the sample mean of \( x \) averaging across stocks. The first column is the correlation with transaction time volatility, the second column with shuffled transaction real time volatility, the third with volume time volatility, and the fourth with shuffled volume real time volatility. The fifth column is the ratio of the first and second columns and the sixth column is the ratio of the third and fourth columns. The fact that the values in the last column are all greater than one shows that fluctuations in transaction number and volume are minority effects on real time volatility.

a series requires sampling the price at intervals of constant volume. This is complicated by the fact that transactions occur at discrete time intervals and have highly variable volume, which causes volume time \( \tau_v(t) \) to increase in discrete and irregular steps. This makes it impossible to sample volume at perfectly regular intervals. We have explored several methods for sampling at approximately regular intervals, and have found that all the sampling procedures we tested yielded essentially the same results. For the results we present here, we chose the simple method of adjusting the sampling interval upward to the next transaction.

To make it completely clear how our results are obtained, we now describe our algorithm for approximate construction of equal volume intervals in more detail. To ensure that the number of intervals in the volume time series is roughly equal to the number in the real time series, we choose a volume time sampling interval \( \Delta V \) so that \( \Delta V \approx \frac{\sum_i V_i}{n} \), where \( n \) is the number of real time intervals and \( \sum_i V_i \) is the total transaction volume for the entire time series. We construct sampling intervals by beginning at the left boundary of each interval. Assume the first transaction in the interval is transaction \( i \). We examine transactions \( k = i - K, \ldots, i \), aggregating the volume \( V_k \) of each transaction and increasing \( K \) incrementally until \( \sum_{k=i-K} V_k > \Delta V \). If the resulting time interval \( (t_{i-K}, t_i) \) crosses a daily boundary we discard it. The volume time volatility is then defined to be \( \nu_v(t_i, \Delta V) = |p(t_i) - p(t_{i-K})| \), i.e., the absolute price change across the smallest interval whose volume is greater than \( \Delta V \). We have also tried other more exact procedures, such as defining the price at the right boundary of the interval by interpolating between \( p(t_{i-K}) \) and \( p(t_{i-K+1}) \), but they give essentially the same results.

In a manner analogous to the definition of shuffled transaction real time volatility, we can also define the shuffled volume real time volatility \( \tilde{\nu}_v \). This is constructed as a point of comparison to match the volume in each interval of
the real time series but otherwise to destroy any time correlations. Our algorithm for computing the shuffled real time volatility is as follows: we begin at a time \( t \) corresponding to the beginning of a real time interval \((t - \Delta t, t)\) that we want to match, which contains transaction volume \( V_{t-\Delta t,t} \). We then randomly sample transactions \( i \) from the entire time series and aggregate their volumes \( V_i \), increasing \( K \) until \( \sum_{i=1}^{K} V_i > V_{t-\Delta t,t} \). We define the aggregate return for this interval as \( \tilde{R}_{t-\Delta t,t} = \sum_{i=1}^{K} \tilde{r}(t_i) \), where \( \tilde{r}(t_i) \) is the (log) return associated with each transaction. The shuffled volume real time volatility is \( \tilde{\nu}_{v}(t, \Delta t) = |\tilde{R}_{t-\Delta t,t}| \).

By comparing the volatility defined in this way to the real time volatility, we can see how important volume is in comparison to other factors. As with transaction time, we can compute the correlations between volume time and real time volatility, \( \rho(\nu, \nu_v) \), and the correlation between shuffled volume real time volatility and real time volatility, \( \rho(\nu, \tilde{\nu}_v) \). The results are summarized by the open marks in Figure 4.3 and columns three, four and six of Table 4.1. In Figure 4.3 we see that \( \rho(\nu, \nu_v) \) is larger than \( \rho(\nu, \tilde{\nu}_v) \) in almost every case. The ratio \( E[\rho(\nu, \nu_v)/\rho(\nu, \tilde{\nu}_v)] \) measures the average relative importance of factors that are independent of volume to factors that depend on volume for each of our three data sets. In column 6 of Table 4.1 these average ratios range from 1.4 - 3.6. As before, other factors have a bigger correlation to volatility than fluctuations in volume. Thus, neither volume fluctuations nor transaction rate fluctuations can account for the majority of variation in volatility.

4.4 Distributional properties of returns

4.4.1 Holding volume or transactions constant

In this section we examine whether fluctuations in either transaction frequency or total trading volume play a major role in determining the distribution of returns. In Figure 4.4 we compare the cumulative distribution of absolute returns in real time to the distribution of absolute returns in transaction time and volume time for a couple of representative stocks\(^5\). For London, where we can separate volume in the on-book and off-book markets, we compare both to total volume time (on-book and off-book) and to volume time based on the on-book volume only. For the NYSE data we cannot separate these, and volume time is by definition total volume time. We see that the differences between real time and either volume or transaction time are not large, indicating that transaction frequency and volume have a relatively small effect on the distribution of prices.

It is difficult to assign error bars to the empirical distributions due to the fact that the data are long-memory (see Section 4.5), so that standard methods tend to seriously underestimate the error bars. However, it is apparent that for the LSE stock Astrazeneca and for the NYSE stock Procter and Gamble in the period where the tick size is a penny, the differences between real time, volume time, and transaction time are minimal. The situation is more complicated for Procter and Gamble in the period where the tick size is 1/8. The distributions

\(^5\)We study absolute returns throughout because the positive and negative tails are essentially the same, and because our primary interest is in volatility.
Figure 4.4: Cumulative distribution of absolute (log) returns $P(|r| > x)$ under several different time clocks, plotted on a double logarithmic scale. (a, top) The LSE stock AZN. (b, middle) The NYSE stock Procter & Gamble (PG) in the 1/8 dollar tick size period. (c, bottom) PG when the tick size is a penny. The time clocks shown are real time (black circles), transaction time (red squares), and volume time (blue diamonds). For comparison we also show a normal distribution (dashed black line) with the same mean and variance as in transaction time. For AZN we also show volume time for the on-book market only (green triangles). The real time returns are 15 minutes in every case; transaction and volume time intervals are chosen to match this. For both LSE and NYSE2 both the transaction time and volume time distributions are almost identical to real time, even in the tails. For PG in the NYSE1 data set the deviations in the tail from real time are noticeable, but they are still well away from the normal distribution.
for volume and transaction time are nearly identical. While they are still clearly heavy tailed, the tails are not as heavy as they are in real time. This leaves open the possibility that during the period of the 1/8 tick size in the NYSE, volume or transaction frequency fluctuations may have played a role in generating heavy tails in the distribution. Nonetheless, even here it is clear that they are not the only factor generating heavy tails.

To see whether these patterns are true on average, in Figure 4.5 we show similar plots, except that here we compute an average of the distributions over each of our data sets. We first normalize the data so that the returns for each stock have the same standard deviation, and then we average the distributions for each stock with equal weighting. For many purposes this can be a dangerous procedure; as we discuss in Section 4.4.3, if different stocks have different tail behaviors this can lead to a biased estimate of the tail exponent. However, in this case, we are just making a qualitative comparison, and we do not think this is a problem.

For the LSE stocks, all the distributions appear to be the same except for statistical fluctuations. The volume time distribution based on on-book volume appears to have somewhat heavier tails than the real time distribution, but for the largest values it joins it again, suggesting that these variations are insignificant. Throughout part of its range the tails for the volume and transaction time distributions for the NYSE1 data set seem to be less heavy, but for the largest fluctuations they rejoin the real time distribution, again suggesting that the differences are statistically insignificant. For NYSE2, transaction time is almost indistinguishable from real time; there is a slight steepening of the tail for volume time, but it would be difficult to argue that there is any statistically significant difference in the tail exponents. These results demonstrate that fluctuations in volume or number of transactions do not have much effect on the return distribution, and are not sufficient to explain its heavy tails.

### 4.4.2 Comparison to work of Ane and Geman

Our results appear to contradict those of Ane and Geman [117]. They studied the hypothesis $Y(t) = X(\tau(t))$, where $Y(t)$ is the stochastic process generating real returns $r(t)$, $X$ is a Brownian motion, and $\tau(t)$ is a stochastic time clock. They developed a method to find the stochastic time clock $\tau$ that would satisfy this assumption, by assuming that the increments $\Delta \tau(t) = \tau(t) - \tau(t - 1)$ of the timeclock are IID and computing the moments of its distribution. They then applied this method to data from two stocks, Cisco and Intel, from January 2 to December 31 in 1997. Using several different time scales ranging from one to fifteen minutes they computed the moments of $\Delta \tau$ needed to make $X$ a Brownian motion, and compared them to the moments of transaction time increments $\Delta \tau_9$ (which is just the number of transactions in a given time). They found that the two sets of moments were very close. This indicates that the non-normal behavior of the distribution of transactions is sufficient to explain the non-normality of stock returns.

Because this conclusion is quite different from ours, we test their hypothesis
Figure 4.5: Cumulative distribution of normalized absolute (log) returns $P(|r| > x)$ averaged over all the stocks in each data set under several different time clocks, plotted on double logarithmic scale. (a, top) LSE, (b, middle) NYSE1, (c, bottom) NYSE2. The time clocks shown are real time (black circles), transaction time (red squares), and volume time (blue diamonds). For the LSE we also show volume time for the on-book market only (green triangles). For comparison we also show a normal distribution (dashed black line) with the same mean and variance as in transaction time.
Figure 4.6: Cumulative distribution of transaction normalized absolute (log) returns \( P(|z(t)|) \), where \( z(t) = y(t)/\sqrt{N(t)} \), for the stocks Cisco (black circles) and Intel (red filled circles) during the period of the Ane and Geman study, Procter & Gamble of NYSE1 (green squares), and of NYSE2 (blue filled squares). These are plotted on double logarithmic scale. For comparison we also show a normal distribution (dashed black line). All four distributions are roughly the same, and none of them is normal. on the same data that they used, attempting to reproduce their procedure for demonstrating normality of returns in transaction time\(^6\). We construct the returns \( r(t) = p(t) - p(t - \Delta t) \) with \( \Delta t = 15 \) minutes, and count the corresponding number of transactions \( N(t) \) in each interval. We then create transaction normalized returns \( z(t) = r(t)/\sqrt{N(t)} \). We construct distributions for Intel and Cisco on the same time periods of their study, and for comparison we also do this for Procter and Gamble over the time periods used in our study for the NYSE1 and NYSE2 data sets. The results are shown in Figure 4.6. In the figure it is clear that all four distributions are quite similar, and that none of them is normal. As a further check we performed the Bera-Jarque test for normal behavior, getting \( p \) values for accepting normality for every stock less than \( 10^{-23} \). We do not know why our results differ from those of Ane and Geman, but we note that Li has concluded that their moment expansion contained a mistake [124].

\(^6\)We are attempting to reproduce Figure 4.4 of reference [117]; it is not completely clear to us what algorithm they used to produce this. The procedure we describe here is our best interpretation, which also seems to match with that of Deo et al. [123].
4.4.3 Implications for the theory of Gabaix et al.

Our results here are relevant to the theory of price formation recently proposed by Gabaix et al. (2003) [75]. They have argued that the price impact of a large volume trade is of the form

\[ r_i = k \epsilon_i V_i^{1/2} + u_i, \]  

(4.1)

where \( r_i \) is the return generated by the \( i^{th} \) trade, \( \epsilon_i \) is its sign (+1 if buyer-initiated and -1 if seller initiated), \( V_i \) is its size, \( k \) is a proportionality constant and \( u_i \) is an uncorrelated noise term. Under the assumption that all the variables are uncorrelated, if the returns are aggregated over any given time period the squared return is of the form

\[ E[r^2|V] = k^2 V + E[u^2], \]  

(4.2)

where

\[ r = \sum_i^N r_i, \quad V = \sum_i^N V_i, \quad E[u^2] = \sum_i^N E[u_i^2], \quad \text{and } N \text{ is the number of transactions, which can vary.} \]

They have hypothesized that equation 4.2 can be used to infer the tail behavior of returns from the distribution of volume. Their earlier empirical work found that the distribution of volume has heavy tails that are asymptotically of the form

\[ P(V > v) \sim v^{-\alpha}, \quad \text{with } \alpha \approx 1.5 \ [99]. \]

These two relations imply that the tails of returns should scale as

\[ P(r > R) \sim R^{-2\alpha}. \]

This theory has been criticized on several grounds. The most important points that have been raised are that Eq. 4.1 is not well supported empirically, that \( \epsilon_i \) are strongly positively correlated with long-memory so that the step from Eq. 4.1 to Eq. 4.2 is not valid, and that at the level of individual transactions \( P(r_i > x|V_i) \) only depends very weakly on \( V_i \) as it is discussed in Chapter 3. There we showed that the returns were independent of volume at the level of individual transactions, and only for the on-book market of the LSE. Here we show this for 15 minute intervals, we control for volume in both the on-book and off-book markets, and we also study NYSE stocks. Our results here provide additional evidence. The distribution of returns in volume time plotted in Figures 4.4 and 4.5 is just the distribution of returns when the volume is held constant, and can be written as \( P(r > x|V = \Delta V) \), where \( \Delta V \) is the sampling interval in volume time as described in Section 4.3.4. We can express their hypothesis in a more general different form by squaring Eq. 4.1 and rewriting Eq. 4.2 in the form

\[ r^2 = k^2 V + w, \]

where

\[ w = r^2 - k^2 V \]

is a noise term that captures all the residuals; this allows for the possibility that \( w \neq u_i^2 \). When \( V \) is held constant \( r \) should be of the form

\[ r = \sqrt{C + w}, \]

where \( C = k^2 V \) is a constant. According to their theory the noise term \( w \) should be unimportant for the tails. Instead, we see that in every case \( P(r > x|V) \) is heavy tailed, and in many cases it is almost indistinguishable from \( P(r > x) \). This indicates that \( w \) is not negligible, but is rather the dominant factor.

This is puzzling because Gabaix et al. have presented empirical evidence that when a sufficient number of stocks are aggregated together \( E[r^2|V] \) satisfies Eq. 4.2 reasonably well. To understand how this can coexist with our results, we make some alternative tests of their theory. We begin by studying \( E[r^2|V] \). Following the procedure they outline in their paper, using 15 minute intervals we
normalize $r$ for each stock by subtracting the mean and dividing by the standard deviation, and normalize $V$ for each stock by dividing by the mean. We allow a generalization of the assumption by letting $E[r^2|V] = k^2V^\beta$, i.e., we do not require that $\beta = 1$. Using least squares regression, for large values of $r$ and $V$ we find $k \approx 0.61$ and $\beta \approx 1.2$. Their theory predicts that $P(r > x) \sim P(kV^{\beta/2} > x)$ for large $x$. In Figure 4.7a we use this together with the empirical distribution of volume $P(V)$ to estimate $P(r > x)$ and compare it to the empirical distribution. Regardless of whether we use $\beta = 1.2$ or $\beta = 1$, the result has a much thinner tail than the actual return distribution\(^7\).

In Figure 4.7b we perform an alternative test. We estimate the predicted return in each 15 minute interval according to their theory as $\tilde{r} = A + kz$, where $z = \sum_i \epsilon_i V_i^{1/2}$, using least squares approximation to estimate $A$ and $k$ in each interval. This can be used to compute a residual $\eta = r - \tilde{r}$. We then compare the distributions of $r$, $\tilde{r}$, and $\eta$ according to their theory $\tilde{r}$ should match the tail behavior and $P(\eta)$ should be unimportant. Instead we find the opposite. $P(\eta)$ is closer to the empirical distribution $P(r)$, roughly matching its tail behavior, while $P(\tilde{r})$ falls off more rapidly – for the largest fluctuations it is about an order of magnitude smaller than the empirical distribution. This indicates that the reason their theory fails is that the heavy-tailed behavior it predicts is dominated by the even heavier-tailed behavior of the residuals $\eta$. In our study of the LSE in Chapter 3 we were able to show explicitly why this is true. There we demonstrated that heavy tails in returns at the level of individual transactions are primarily caused by the distributions of gaps (regions without any orders) in the limit order book, whose size is unrelated to trading volume.

### 4.5 Long-memory

There is mounting evidence that volatility is a long-memory process [128, 129, 130, 131]. A long-memory process is a random process with an autocorrelation function that is not integrable. This typically happens when the autocorrelation function decays asymptotically as a power law of the form $\tau^{-\gamma}$ with $\gamma < 1$. The existence of long-memory is important because it implies that values from the distant past have a significant effect on the present and that the stochastic process lacks a typical time scale. A stochastic process that is built out of a sum whose increments have long-memory has anomalous diffusion, i.e., the variance grows as $\tau^{2H}$, where $H > 0.5$, and $H$ is the Hurst exponent, which is defined below. Statistical averages of long-memory processes converge slowly.
Figure 4.7: Test of the theory of Gabaix et al. In (a) we compare the empirical return distribution to the prediction of the theory of Gabaix et al. for the NYSE1 data set. The solid black curve is the empirical distribution $P(r > x)$, the dashed red curve is the distribution of predictions $\hat{r}$ under the assumption that $\beta = 1.2$, and the blue curve with circles is the same thing with $\beta = 1$. In (b) we compare the empirical distribution $P(r > x)$ (black), the distribution of residuals $P(\eta > x)$ (blue squares), and the predicted return under the theory of Gabaix et al., $P(\hat{r} > x)$ (red circles). The dashed black line is a normal distribution included for comparison.
Plerou et al. [96] demonstrated that the number of transactions in NYSE stocks is a long-memory process. This led them to conjecture that fluctuations in the number of transactions are the proximate cause of long-memory in volatility. This hypothesis makes sense in light of the fact that either long-memory or sufficiently fat tails in fluctuations in transaction frequencies are sufficient to create long-memory in volatility [123]. However, our results so far suggest caution – just because it is possible to create long-memory with this mechanism does not mean it is the dominant mechanism – there may be multiple sources of long-memory.

If there are two long memory processes with autocorrelation exponents $\gamma_1$ and $\gamma_2$, if $\gamma_1 < \gamma_2$ the long-memory of process one will dominate that of process two, since for large values of $\tau$, $C_1(\tau) \gg C_2(\tau)$. We will see that this is the case here – the long-memory caused by fluctuations in volume and number of transactions is dominated by long-memory caused by other factors.

We investigate the long-memory of volatility by computing the Hurst exponent $H$. For a random process whose second moment is finite and whose autocorrelation function asymptotically decays as a power law $\tau^{-\gamma}$ with $0 < \gamma < 1$, the Hurst exponent is $H = 1 - \gamma/2$. A process has long memory if $1/2 < H < 1$. We compute the Hurst exponent rather than working directly with the autocorrelation function because it is better behaved statistically. We compute the Hurst exponent using the DFA method [132]. The time series $\{x(t)\}$ is first integrated to give $X(t) = \sum_{i=1}^{t} x(i)$. For a data set with $n$ points, $X(t)$ is then broken into $K$ disjoint sets $k = 1, \ldots, K$ of size $L \approx n/K$. A polynomial $Y_{k,L}$ of order $m$ is then fit for each set $k$ using a least squares regression. For a given value of $L$ let $D(L)$ be the root mean square deviation of the data from the polynomials averaged over the $K$ sets, i.e., $D(L) = \frac{1}{K} \sum_{k,i} (X(i) - Y_{k,L}(x(i)))^2$. The Hurst exponent is formally defined as $H = \lim_{L \to \infty} \lim_{n \to \infty} \log D(L) / \log L$. In practice, with a data set of finite $n$ the Hurst exponent is measured by regressing $\log D(L)$ against $\log L$ over a region with $L_{\text{min}} < L < n/4$. $H$ is the slope of the regression.

To test the hypothesis that the number of transactions drives the long-memory of volatility, we compute the Hurst exponent of the real-time volatility, $H(\nu)$, and compare it to the Hurst exponent of the volatility in transaction time, $H(\nu_0)$, and the Hurst exponent in shuffled transaction real time, $H(\tilde{\nu}_0)$. If the long-memory of transaction time is the dominant cause of volatility, then we should see that $H(\nu) \approx H(\tilde{\nu}_0)$, due to the fact that $\tilde{\nu}$ preserves the number of transactions in each 15 minute interval, and we should also see that $H(\nu) > H(\nu_0)$, due to the fact that $\nu_0$ holds the number of transactions constant in each interval, so it should not display as much clustered volatility. We illustrate the scaling diagrams used to compute these three Hurst exponents for the LSE stock Astrazeneca in Figure 4.8. This figure does not support the conclusion that transaction time fluctuations are the proximate cause of volatility. We find $H(\nu) \approx 0.70 \pm 0.07$, $H(\nu_0) \approx 0.70 \pm 0.07$, and $H(\tilde{\nu}_0) \approx 0.59 \pm 0.03$. Thus, for Astrazeneca it seems the opposite is true, i.e., $H(\nu) \approx H(\nu_0)$ and $H(\nu) > H(\tilde{\nu}_0)$. While it is true that $H(\tilde{\nu}_0) > 0.5$, which means that fluctuations in transaction frequency contribute to long-memory, the fact that $H(\nu_0) > H(\tilde{\nu}_0)$ means that this is dominated by even stronger long-memory effects that are caused by other factors.
Figure 4.8: Computation of Hurst exponent of volatility for the LSE stock Astrazeneca. The logarithm of the average variance $D(L)$ is plotted against the scale $\log L$. The Hurst exponent is the slope. This is done for real-time volatility $\nu$ (black circles), transaction time volatility $\nu_{\theta}$ (red crosses), and shuffled transaction real-time volatility $\tilde{\nu}_{\theta}$ (blue triangles). The slopes in real time and transaction time are essentially the same, but the slope for shuffled transaction real time is lower, implying that transaction fluctuations are not the dominant cause of long-memory in volatility.
Figure 4.9: Hurst exponents for alternative volatilities vs. real-time volatility. Each point corresponds to a particular stock and alternative volatility measure. NYSE1 stocks are in red, LSE in green, and NYSE2 in blue. The solid circles are transaction time, open circles are shuffled transaction real time, solid diamonds are volume time, and open diamonds are shuffled volume real time. The black dashed line is the identity line. The fact that transaction-time and volume-time Hurst exponents cluster along the identity line, whereas almost all of the shuffled real time values are well away from it, shows that neither volume nor transaction fluctuations are dominant causes of long-memory.

Note that the quoted error bars for $H$ as based on the assumption that the data are normal and IID, and are thus much too optimistic. For a long-memory process such as this the relative error scales as $n^{(H-1)}$ rather than $n^{-1/2}$, and estimating accurate error bars is difficult. The only known procedure is the variance plot method [133], which is not very reliable and is tedious to implement. This drives us to make a cross-sectional test, where the consistency of the results and their dependence on other factors will make the statistical significance quite clear.

To test the consistency of our results above, we compute Hurst exponents for all the stocks in each of our three data sets. In addition, to test whether fluctuations in volume are important, we also compute the Hurst exponents of volatility in volume time, $H(\nu_v)$ and in shuffled volume real time, $H(\tilde{\nu}_v)$. The results are shown in Figure 4.9, where we plot the Hurst exponents for $H(\nu_\theta)$, $H(\tilde{\nu}_\theta)$, $H(\nu_o)$, and $H(\tilde{\nu}_o)$ against the real-time volatility $H(\nu)$ for each stock in each data set. Whereas the Hurst exponents in volume and transaction time cluster along the identity line, the Hurst exponents for shuffled real time are
further away from the identity line and are consistently lower in value. This is seen at a glance in the figure by the fact that the solid marks are clustered along the identity line whereas the open marks are scattered away from it. The results are quite consistent – out of the 60 cases shown in Figure 4.9, only one has $H(\tilde{\nu}_v) > H(\nu_v)$ and only one has $H(\tilde{\nu}_\theta) > H(\nu_\theta)$.

As a further test we perform a regression of the form $H_i^{(a)} = a + bH_i^{(r)}$, where $H_i^{(a)}$ is an alternative Hurst exponent and $H_i^{(r)}$ is the Hurst exponent for real-time volatility for the $i^{th}$ stock. We do this for each data set and each alternative volatility measure. The results are presented in Table 4.2. For volume and transaction time all of the slope coefficients $b$ are positive, all but one at the 95% confidence level. This shows that as the long-memory of real-time volatility increases, the volatility when the volume or number of transactions is held constant increases with it. In contrast, for the shuffled volume real-time measures, four out of six cases have negative slopes and none of them is statistically significant. This shows that the long-memory of real-time volatility is not driven by the volume or the transaction frequency. The order of events is important – more important than their number. The $R^2$ values of the regressions are all higher for volume or transaction time than for their corresponding shuffled values, and the mean distances from the identity line are substantially higher. These facts taken together make it clear that the persistence of real-time volatility is driven more strongly by other factors than by volume or transaction fluctuations.

It is interesting that the data set appears to be the most significant factor determining $H$. For the NYSE the real-time Hurst exponents during the 1/8 tick size period are all in the range $0.62 < H < 0.73$, whereas in the penny tick size period they are in the range $0.77 < H < 0.83$. Thus, the Hurst exponents for the two periods are completely disjointed – all the values of $H$ during the penny tick size period are higher than any of the values during the 1/8 tick size period. The LSE Hurst exponents are roughly in the middle, spanning the range $0.72 < H < 0.82$. It is beyond the scope of this study to determine why this is true, but this suggests that changes in tick size or other aspects of market structure are very important in determining the strength of the persistence of volatility.
Table 4.2: A summary of results comparing Hurst exponents for alternative volatilities to real-time volatility. We perform regressions on the results in Figure 4.9 of the form \( H_i^{(a)} = a + b H_i^{(r)} \), where \( H_i^{(a)} \) is an alternative Hurst exponent and \( H_i^{(r)} \) is the Hurst exponent for real-time volatility for the \( i \)th stock. We do this for each data set and each of the exponents \( H(\nu_\theta) \) (transaction time), \( H(\tilde{\nu}_\theta) \) (shuffled transaction real time), \( H(\nu_v) \) (volume time), and \( H(\tilde{\nu}_v) \) (shuffled volume real time). \( d \) is the average distance to the identity line \( H_i^{(a)} = H_i^{(r)} \) and \( R^2 \) is the goodness of fit of the regression. We see that in the top two rows \( b \) is positive and statistically significant in all but one case, in contrast to the bottom two rows. This and the fact that \( d \) is much smaller in the top two rows makes it clear that neither volume nor transactions are important causes of long-memory.

<table>
<thead>
<tr>
<th>volatility</th>
<th>market</th>
<th>( a )</th>
<th>( b )</th>
<th>( d \times 100 )</th>
<th>( R^2 \times 100 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nu_\theta )</td>
<td>NYSE1</td>
<td>0.15 ( \pm ) 0.10</td>
<td>0.76 ( \pm ) 0.20</td>
<td>1.4</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>NYSE2</td>
<td>-0.27 ( \pm ) 0.11</td>
<td>1.34 ( \pm ) 0.14</td>
<td>0.76</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>LSE</td>
<td>0.26 ( \pm ) 0.11</td>
<td>0.65 ( \pm ) 0.15</td>
<td>1.1</td>
<td>52</td>
</tr>
<tr>
<td>( \nu_v )</td>
<td>NYSE1</td>
<td>0.46 ( \pm ) 0.12</td>
<td>0.35 ( \pm ) 0.19</td>
<td>1.4</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>NYSE2</td>
<td>0.17 ( \pm ) 0.23</td>
<td>0.75 ( \pm ) 0.29</td>
<td>2.0</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>LSE</td>
<td>0.25 ( \pm ) 0.16</td>
<td>0.67 ( \pm ) 0.21</td>
<td>1.6</td>
<td>37</td>
</tr>
<tr>
<td>( \tilde{\nu}_\theta )</td>
<td>NYSE1</td>
<td>0.65 ( \pm ) 0.18</td>
<td>-0.03 ( \pm ) 0.25</td>
<td>4.0</td>
<td>8.3 ( \times ) 10^{-2}</td>
</tr>
<tr>
<td></td>
<td>NYSE2</td>
<td>0.59 ( \pm ) 0.29</td>
<td>0.03 ( \pm ) 0.35</td>
<td>13</td>
<td>4.6 ( \times ) 10^{-2}</td>
</tr>
<tr>
<td></td>
<td>LSE</td>
<td>0.47 ( \pm ) 0.12</td>
<td>0.19 ( \pm ) 0.16</td>
<td>10</td>
<td>8.1</td>
</tr>
<tr>
<td>( \tilde{\nu}_v )</td>
<td>NYSE1</td>
<td>0.67 ( \pm ) 0.18</td>
<td>-0.04 ( \pm ) 0.26</td>
<td>3.2</td>
<td>1.2 ( \times ) 10^{-1}</td>
</tr>
<tr>
<td></td>
<td>NYSE2</td>
<td>1.11 ( \pm ) 0.25</td>
<td>-0.52 ( \pm ) 0.31</td>
<td>7.3</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>LSE</td>
<td>0.64 ( \pm ) 0.11</td>
<td>-0.18 ( \pm ) 0.15</td>
<td>18</td>
<td>7.5</td>
</tr>
</tbody>
</table>
Chapter 5

Conclusions

In this study we analyze limit order based trading by introducing and simulating a simplistic model and accomplishing real data analysis mostly based on the well detailed data of the London Stock Exchange in order to gain a better understanding of the characteristics and the causes of price diffusion.

First, in the spirit of Gode and Sunder [30], we assumed a simple, zero-intelligence model of agent behavior and show that the trading mechanism exerts considerable power in shaping the properties of prices. While not disputing that agent behavior might be important, our model suggests that, at least on the short timescale, many of the properties of the market are dictated by the trading mechanism (market institution), and in particular the need to store supply and demand. Our model is stochastic and fully dynamic, and makes predictions that go beyond the realm of experimental economics, giving quantitative predictions about fundamental properties of a real market. We have developed what were previously conceptual toy models in the physics literature into a model with testable explanatory power.

This raises questions about the comparison to standard models based on the response of valuations to news. The idea that news might drive changes in order flow rates is compatible with our model. That is, news can drive changes in order flow, which in turn cause the best bid or ask price to change. But notice that in our model there are no assumptions about valuations. Instead, everything depends on order flow rates. For example, the diffusion rate of prices increases as the 5/2 power of market order flow rate, and thus volatility, which depends on the square root of the diffusion rate, increases as the 5/4 power. Of course, order flow rates can respond to information; an increase in market order rate indicates added impatience, which might be driven by changes in valuation. But changes in long-term valuation could equally well cause an increase in limit order flow rate, which decreases volatility. Valuation per se does not determine whether volatility will increase or decrease. Our model says that volatility does not depend directly on valuations, but rather on the urgency with which they are felt, and the need for immediacy in responding to them.

Understanding the shape of the price impact function was one of the motivations that originally set this modeling study into motion. The price impact function is closely related to supply and demand functions, which have been cen-
tral aspects of economic theory since the 19th century. Our model suggests that
the shape of price impact functions in modern markets is significantly influenced
not so much by strategic thinking as by an economic fundamental: the need to
store supply and demand in order to provide liquidity. \textit{A priori} it is surprising
that this requirement alone may be sufficient to dictate at least the broad outlines
of the price impact curve.

This simplistic model offers a “divide and conquer” strategy to understanding
fundamental problems in economics. Rather than trying to ground our approach
directly on assumptions of utility, we break the problem into two parts. We
provide an understanding of how the statistical properties of prices respond to
order flow rates, and leave the problem open of how order flow rates depend on
more fundamental assumptions about information and utility. Order flow rates
have the significant advantage that, unlike information, utility, or the cognitive
powers of an agent, they are directly measurable. We believe that by breaking the
problem into two pieces, and partially solving the second piece, we can ultimately
help provide a deeper understanding of how markets work.

This view gained confirmation when we studied the cause of large individ-
ual fluctuations in prices in the London Stock Exchange. This was done at the
microscopic level of individual events, where an event is the placement or can-
cellation of an order to buy or sell. We showed that price fluctuations caused
by individual market orders are essentially independent of the volume of orders.
Instead, large price fluctuations are driven by liquidity fluctuations, variations in
the market’s ability to absorb new orders. Even for the most liquid stocks there
can be substantial gaps in the order book, corresponding to a block of adjacent
price levels containing no quotes. When such a gap exists next to the best price,
a new order can remove the best quote, triggering a large midpoint price change.
Thus, the distribution of large price changes merely reflects the distribution of
gaps in the limit order book. This is a finite size effect, caused by the granularity
of order flow: in a market where participants placed many small orders uniformly
across prices, such large price fluctuations would not happen. We present results
suggesting that the risk profile varies from stock to stock, and is not universal:
lightly traded stocks tend to have more extreme risks.

At a higher level, these results demonstrate that large price changes are driven
by fluctuations in liquidity. There are times when the market absorbs changes
in supply and demand smoothly, and other times when a small change in supply
or demand can result in a very large change in the price. This is due to the
fact that supply and demand functions are not smooth, but rather have large,
irregular steps and jumps. The market is granular, due to the presence of only a
finite number of occupied price levels in the book. Even for an active stock such
as AZN, the number of occupied price levels on one side of the book at any given
time is typically about 30, and so the system is far from the continuum limit.
However, we have shown that this is not a simple matter of fluctuations in the
number of occupied price levels; while there is an approximate power law in the
limit $N \to 0$ in the frequency for occupying $N$ price levels, this is not sufficient to
explain fluctuations in prices. Instead, the power law in gaps persists even when
the number of occupied levels is quite high, reflecting nontrivial correlations in
the positions of orders sitting in the book.

The work presented here suggests that it is important to properly model trading mechanisms. However, at this point it is not clear how these results would change for a substantially different market institution. Fat tails are also observed in markets, such as the London Metals Exchange, that follow very different exchange mechanisms. While the LSE follows a continuous double auction, it is our belief that key elements are likely to persist with other market mechanisms. In particular, we hypothesize that large price fluctuations in any market are driven by liquidity fluctuations, and that the granularity of fluctuations in supply and demand remains the key factor underlying extreme price fluctuations.

In order to examine the relevance of elementary trading mechanisms for longer time scales, we analyzed longer scale returns using different time measures. The idea that clustered volatility and heavy tails in price returns can be explained by fluctuations in transactions or volume is seductive in its simplicity. Both transaction frequency and volume are strongly positively correlated with volatility, and it is clear that fluctuations in volume (or transaction frequency) can cause both clustered volatility and heavy tails, so this might seem to be an open and shut case. Our main result is to show that this is not true, at least for the data that we have studied here. For these data the effects of volume and transaction frequency are dominated by other effects. We have shown this for three different properties, the contemporaneous relationship with the size of price changes, the long-memory properties of volatility, and the distribution of returns. The results have been verified with three different large data sets, with tens of millions of transactions, and changes in both the time period and the market. The story is quite consistent in each case, with only a few minor variations.

This can be viewed as a tale of competing effects. We do not dispute that volume and transaction frequency affect prices, but for these data other effects are more important. If there were no other competing effects, fluctuations in volume and transaction would indeed cause clustered volatility and heavy tails. However, in this case the clustered volatility would not be as strong, it would not be as persistent, and there would be fewer extreme price returns.

The most interesting aspect of our results is their implication for price formation. If volume and transaction frequency are not the primary proximate causes of price movements, then what is? It is surprising that after so many years of empirical work the answer to this question remains unknown. This study on volatility we have done so far supports the main themes of our study about large price changes. There we showed that for the LSE the heavy tails in prices are already present at the microscopic level of individual price fluctuations. The tails of the distribution are essentially unaltered in transaction time, and the size of individual transactions has only a small effect on the size of price changes. Taking advantage of the precision of the LSE data set, we looked at midprice changes caused by limit orders placed inside the spread, market orders that remove all the depth at the best prices, and cancellations of the last order at the best price. We found that they all have essentially the same distribution. We attribute this fact to the dynamics of gaps in the limit order book. As price changes are aggregated on longer time scales, the heavy-tailed behavior on long time scales mimics that
on shorter time scales, even if the crossover from the body to heavy-tailed behavior occurs at successively larger values. These results indicate that the main driver of heavy tails is the heavy-tailed distribution of individual price changes themselves. This is driven by order placement and cancellation, and not by volume or transaction frequency. This is by no means an ultimate cause, but it does provide insight into where such a cause might lie.

This story for the origin of heavy tails is consistent with our work so far studying clustered volatility. We have shown here that the ordering of transactions is very important in determining clustered volatility. If we scramble the order of transactions but preserve their frequency, clustered volatility is greatly diminished. A similar statement applies to volume, making it clear that the size of transactions is also not playing an important role.
Appendix A

Theoretical analysis of the simplistic model

A.1 Summary of analytic methods

We have investigated this model analytically using two approaches. The first one is based on a master equation, given in Section A.6. This approach works best in the midpoint centered frame. Here we attempt to solve directly for the average number of shares at each price tick as a function of price. The midpoint price makes a random walk with a nonstationary distribution. Thus, the key to finding a stationary analytic solution for the average depth is to use comoving price coordinates, which are centered on a reference point near the center of the book, such as the midpoint or the best bid. In the first approximation, fluctuations about the mean depth at adjacent prices are treated as independent. This allows us to replace the distribution over depth profiles with a simpler probability density over occupation numbers $n$ at each $p$ and $t$. We can take a continuum limit by letting the tick size $dp$ become infinitesimal. With finite order flow rates, this gives vanishing probability for the existence of more than one order at any tick as $dp \to 0$. This is described in detail in Section A.6.3. With this approach, we are able to test the relevance of correlations as a function of the parameter $\epsilon$ as well as predict the functional dependence of the cumulative distribution of the spread on the depth profile. It is seen that correlations are negligible for large values of $\epsilon (\epsilon \sim 0.2)$ while they are very important for small values ($\epsilon \sim 0.002$).

Our second analytic approach, which we term the Independent Interval Approximation (IIA), is most easily carried out in the bid-centered frame and is described in Section A.7. This approach uses a different representation, in which the solution is expressed in terms of the empty intervals between non-empty price ticks. The system is characterized at any instant of time by a set of intervals \{...$x_{-1}$, $x_0$, $x_1$, $x_2$...\} where, for example, $x_0$ is the distance between the bid and the ask (the spread), $x_{-1}$ is the distance between the second buy limit order and the bid, and so on (see Figure A.1). Equations are written for how a given interval varies in time. Changes to adjacent intervals are related, giving us an infinite set of coupled non-linear equations. However, using a mean-field approximation
we are able to solve the equations, albeit only numerically. Besides predicting how the various intervals (for example the spread) vary with the parameters, this approach also predicts the depth profiles as a function of the parameters. The predictions from the IIA are compared to data from numerical simulations, in Section A.7.2. They match very well for large \( \epsilon \) and less well for smaller values of \( \epsilon \). The IIA can also be modified to incorporate various extensions to the model, as mentioned in Section A.7.2.

In both approaches, we use a mean field approximation to get a solution. The approximation basically lies in assuming that fluctuations in adjacent intervals (which might be adjacent price ranges in the master equation approach or adjacent empty intervals in the IIA) are independent. Also, both approaches are most easily tractable only in the continuum limit \( dp \to 0 \), when every tick has at most only one order. They may, however, be extended to general tick size as well. This is explained in the appendix for the Master Equation approach.

Because correlations are important for small \( \epsilon \), both methods work well mostly in the large \( \epsilon \) limit, though qualitative aspects of small \( \epsilon \) behavior may also be gleaned from them. Unfortunately, at least based on our preliminary investigation of London Stock Exchange data, it seems that it is this small \( \epsilon \) limit that real markets may tend more towards. So our approximate solutions may not be as useful as we would like. Nonetheless, they do provide some conceptual insights into what determines depth and price impact.

In particular, we find that the shape of the mean depth profile depends on a single parameter \( \epsilon \), and that the relative sizes of its first few derivatives account for both the order size-dependence of the market impact, and the renormalization of the midpoint diffusivity. A higher relative rate of market versus limit orders depletes the center of the book, though less than the classical estimate predicts. This leads to a more concave impact (explaining Figure 2.8) and faster short-term diffusivity. However, the orders pile up more quickly (versus classically nondimensionalized price) with distance from the midpoint, causing the rapid early diffusion to suffer larger mean reversion. These are the effects shown in Figure 2.11. We will elaborate on the above remarks in the following sections, however, the qualitative relation of impact to midpoint autocorrelation supplies a potential interpretation of data, which may be more robust than details of the model assumptions or its quantitative results.

Both of the treatments described above are approximations. We can derive an exact global conservation law of order placement and removal whose consequences we elaborate in section A.3. This conservation law must be respected in any sensible analysis of the model, giving us a check on the approximations. It also provides some insight into the anomalous diffusion properties of this model.

A.2 Characterizing limit-order books: dual coordinates

We begin with the assumption of a price space. Price is a dimensional quantity, and the space is divided into bins of length \( dp \) representing the ticks, which may be
Figure A.1: The price space and order profile. $n(p,t)$ has been chosen to be 0 or \pm 1, a restriction that will be convenient later. Price bins are labeled by their lower boundary price, and intervals $x(N)$ will be defined below.

Finite or infinitesimal. Prices are then discrete or continuous-valued, respectively.

Statistical properties of interest are computed from temporal sequences or ensembles of limit-order book configurations. If $n$ is the variable used to denote the number of shares from limit orders in some bin $(p,p+dp)$ at the beginning $t$ of an elementary time interval, a configuration is specified by a function $n(p,t)$. It is convenient to take $n$ positive for sell limit orders, and negative for buy limit orders. Because the model dynamics preclude crossing limit orders, there is in general a highest instantaneous buy limit-order price, called the bid $b(t)$, and a lowest sell limit-order price, the ask $a(t)$, with $b(t) < a(t)$ always. The midpoint price, defined as $m(t) \equiv \frac{a(t) + b(t)}{2}$, may or may not be the price of any actual bin, if prices are discrete ($m(t)$ may be a half-integer multiple of $dp$).

These quantities are diagrammed in Figure A.1.

An equivalent specification of a limit-order book configuration is given by the cumulative order count

$$N(p,t) \equiv \sum_{-\infty}^{p-dp} |n(p,t)| - \sum_{-\infty}^{a-dp} |n(p,t)|,$$

where $-\infty$ denotes the lower boundary of the price space, whose exact value must not affect the results. (Because by definition there are no orders between the bid and ask, the bid could equivalently have been used as the origin of summation. Because price bins will be indexed here by their lower boundaries, though, it is convenient here to use the ask.) The absolute values have been placed so that $N$, like $n$, is negative in the range of buy orders and positive in the range of sells. The construction of $N(p,t)$ is diagrammed in Figure A.2.

In many cases of either sparse orders or infinitesimal $dp$, with fixed order size (which we may as well define to be one share) there will be either zero or one share in any single bin, and Eq. (A.1) will be invertible to an equivalent specification of the limit-order book configuration

$$p(N,t) \equiv \max \{p \mid N(p,t) = N\},$$

shown in Figure A.3. (Strictly speaking, the inversion may be performed for any distribution of order sizes, but the resulting function is intrinsically discrete, so its domain is only invariant when order size is fixed. To give $p(N,t)$ the convenient
Figure A.2: The accumulated order number $N(p, t)$. $N(a, t) \equiv 0$, because contributions from all bins cancel in the two sums. $N$ remains zero down to $b(t) + dp$, because there are no uncanceled, nonzero terms. $N(b, t)$ becomes negative, because the second sum in Eq. (A.1) now contains $n(b, t)$, not canceled by the first.

properties of a well-defined function on an invariant domain, this will be assumed below.)

With definition (A.2), $p(0, t) \equiv a(t)$, $p(-1, t) \equiv b(t)$, and one can define the intervals between orders as

$$x(N, t) \equiv p(N, t) - p(N - 1, t).$$

(A.3)

Thus, $x(0, t) = a(t) - b(t)$, the instantaneous bid-ask spread. The lowest values of $x(N, t)$ bracketing the spread are shown in Figure A.1. For symmetric order-placement rules, probability distributions over configurations will be symmetric under either $n(p, t) \rightarrow -n(-p, t)$, or $x(N, t) \rightarrow x(-N, t)$. Coordinates $N$ and $p$ furnish a dual description of configurations, and $n$ and $x$ are their associated differences. The Master Equation approach of Section A.6 assumes independent fluctuation in $n$ while the Independent Interval Approximation of Sec. A.7 assumes independent fluctuation in $x$. (In this section, it will be convenient to abbreviate $x(N, t) \equiv x_N(t)$.)

### A.3 Frames and marginals

The $x(N, t)$ specification of limit-order book configurations has the property that its distribution is stationary under the dynamics considered here. The same is not true for $p(N, t)$ or $n(p, t)$ directly, because bid, midpoint, and ask prices undergo a random walk, with a renormalized diffusion coefficient. Stationary distributions for $n$-variables can be obtained in *co-moving frames*, of which there are several natural choices.

The *bid-centered configuration* is defined as

$$n_b(p, t) \equiv n(p - b(t), t).$$

(A.4)
Figure A.3: The inverse function $p(N, t)$. The function is in general defined only on discrete values of $N$, so this domain is only invariant when order size is fixed, a convenience that will be assumed below. Between the discrete domain, and the definition of $p$ as a maximum, the inverse function effectively interpolates between vertices of the reflected image of $N(p, t)$, as shown by the dotted line.

If an appropriate rounding convention is adopted in the case of discrete prices, a \textit{midpoint-centered configuration} can also be defined, as

$$n_m(p, t) \equiv n(p - m(t), t).$$  \hfill (A.5)

The midpoint-centered configuration has qualitative differences from the bid-centered configuration, which will be explored below. Both give useful insights into the order distribution and diffusion processes. The ask-centered configuration, $n_a(p, t)$, need not be considered if order placement and removal are symmetric, because it is a mirror image of $n_b(p, t)$.

The \textit{spread} is defined as the difference $s(t) \equiv a(t) - b(t)$, and is the value of the ask in bid-centered coordinates. In midpoint-centered coordinates, the ask appears at $s(t)/2$.

The configurations $n_b$ and $n_m$ are dynamically correlated over short time intervals, but evolve ergodically in periods longer than finite characteristic correlation times. Marginal probability distributions for these can, therefore, be computed as time averages, either as functions on the whole price space, or as discrete sets of prices. Their marginal mean values at a single price $p$ will be denoted $\langle n_b(p) \rangle$, $\langle n_m(p) \rangle$, respectively.

These means are subject to global balance constraints, between total order placement and removal in the price space. Because all limit orders are placed above the bid, the bid-centered configuration obeys a simple balance relation:

$$\frac{\mu}{2} = \sum_{p=b+dp}^{\infty} (\alpha - \delta \langle n_b(p) \rangle).$$  \hfill (A.6)

Eq. (A.6) says that buy market orders must account, on average, for the difference between all limit orders placed, and all decays. After passing to nondimensional
coordinates below, this will imply an inverse relation between corrections to the classical estimate for diffusivity at early and late times, discussed in Sec. A.5. In addition, this conservation law plays an important role in the analysis and determination of the $x(N, t)$’s, as we will see later in the text.

The midpoint-centered averages satisfy a different constraint:

$$\frac{\mu}{2} = \alpha \langle s \rangle + \sum_{p=b+dp}^{\infty} (\alpha - \delta \langle n_m (p) \rangle).$$  \hspace{1cm} (A.7)

Market orders in Eq. (A.7) account not only for the excess of limit order placement over evaporation at prices above the midpoint, but also the “excess” orders placed between $b(t)$ and $m(t)$. Since these always lead to midpoint shifts, they ultimately appear at positive comoving coordinates, altering the shape of $\langle n_m (p) \rangle$ relative to $\langle n_b (p) \rangle$. Their rate of arrival is $\alpha \langle m - b \rangle = \alpha \langle s \rangle /2$. These results are also confirmed in simulations.

### A.4 Factorization tests

Whether in the bid-centered frame or the midpoint centered frame, the probability distribution function for the entire configuration $n(p)$ is too difficult a problem to solve in its entirety. However, an approximate master equation can be formed for $n$ independently at each $p$ if all joint probabilities factor into independent marginals, as

$$\text{Pr} (\{n(p_i)\}_i) = \prod_i \text{Pr} (n(p_i)), \hspace{1cm} (A.8)$$

where Pr denotes, for instance, a probability density for $n$ orders in some interval around $p$.

Whenever orders are sufficiently sparse that the expected number in any price bin is simply the probability that the bin is occupied (up to a constant of proportionality), the independence assumption implies a relation between the cumulative distribution for the spread of the ask and the mean density profile. In units where the order size is one, the relation is

$$\text{Pr} (s/2 < p) = 1 - \exp \left( - \sum_{p'=b+dp}^{p-dp} \langle n_m (p') \rangle \right). \hspace{1cm} (A.9)$$

This relation is tested against simulation results in Figure A.4. One can observe that there are three regimes.

A high-$\epsilon$ regime is defined when the mean density profile at the midpoint $\langle n_m (0) \rangle \lesssim 1$, and strongly concave downward. In this regime, the approximation of independent fluctuations is excellent, and a master equation treatment is expected to be useful. Intermediate-$\epsilon$ is defined by $\langle n_m (0) \rangle \ll 1$ and nearly linear, and the approximation of independence is marginal. Large-$\epsilon$ is defined by $\langle n_m (0) \rangle \ll 1$ and concave upward, and the approximation of independent fluctuations is completely invalid. These regimes of validity correspond also to the qualitative ranges noted already in Sec. 2.2.2.
In the bid-centered frame, however, Eq. A.9 never seems to be valid for any range of parameters. We will discuss later why this might be so. For the present, therefore, the master equation approach is carried out in the midpoint-centered frame. Alternatively, the mean field theory of the separations is most convenient in the bid-centered frame, so that frame will be studied in the dual basis. The relationship of results in the two frames, and via the two methods of treatment, will provide a good qualitative, and for some properties quantitative, understanding of the depth profile and its effect on impacts.

It is possible in a modified treatment, to match certain features of simulations at any $\epsilon$, by limited incorporation of correlated fluctuations. However, the general master equation will be developed independent of these, and tested against simulation results at large $\epsilon$, where its defining assumptions are well met.

### A.5 Comments on renormalized diffusion

A qualitative understanding of why the diffusivity is different over short and long times scales, as well as why it may depend on $\epsilon$, may be gleaned from the following observations.

First, global order conservation places a strong constraint on the classically nondimensionalized density profile in the bid-centered frame. We have seen that at $\epsilon \ll 1$, the density profile becomes concave upward near the bid, accounting for an increasing fraction of the allowed “remainder area” as $\epsilon \rightarrow 0$ (see Figures 2.3 and A.14). Since this remainder area is fixed at unity, it can be conserved only if the density profile approaches one more quickly with increasing price. Low density at low price appears to lead to more frequent persistent steps in the effective short-term random walk, and hence large short-term diffusivity. However, increased density far from the bid indicates less impact from market orders relative to the relaxation time of the Poisson distribution, and thus a lower long-time diffusivity.

The qualitative behavior of the bid-centered density profile is the same as that of the midpoint-centered profile, and this is expected because the spread distribution is stationary, rather than diffusive. In other words, the only way the diffusion of the bid or ask can differ from that of the midpoint is for the spread to either increase or decrease for several succeeding steps. Such autocorrelation of the spread cannot accumulate with time if the spread itself is to have a stationary distribution. Thus, the shift in the midpoint over some time interval can only differ from that of the bid or ask by at most a constant, as a result of a few correlated changes in the spread. This difference cannot grow with time, however, and so does not affect the diffusivity at long times.

Indeed, both of the predicted corrections to the classical estimate for diffusivity are seen in simulation results for midpoint diffusion. The simulation results, however, show that the implied autocorrelations change the diffusivity by factors of $\sqrt{\epsilon}$, suggesting that these corrections require a more subtle derivation than the one attempted here. This will be evidenced by the difficulty of obtaining a source term $S$ in density coordinates (section A.6), which satisfied both the global order conservation law, and the proper zero-price boundary condition, in
the midpoint-centered frame.

An interesting speculation is that the subtlety of these correlations also causes the density $n(p,t)$ in bid-centered coordinates not to approximate the mean-field condition at any of the parameters studied here, as noted in Sec. A.4. Since short-term and long-term diffusivity corrections are related by a hard constraint, the difficulty of producing the late-time density profile should match that of producing the early-time profile. The midpoint-centered profile is potentially easier, in that the late-time complexity must be matched by a combination of the early-time density profile and the scaling of the expected spread. It appears that the complex scaling is absorbed in the spread, as per Figure 2.10 and Figure A.10, leaving a density that can be approximately calculated with the methods used here.

A.6 Master equations and mean-field approximations

There are two natural limits in which functional configurations may become simple enough to be tractable probabilistically, with analytic methods. They correspond to mean field theories in which fluctuations of the dual differentials of either $N(p,t)$ or $p(N,t)$ are independent. In the first case, probabilities may be defined for any density $n(p,t)$ independently at each $p$, and in the second for the separation intervals $x(N,t)$ at each $N$. The mean field theory from the first approximation will be solved in Subsec. A.6.1, and that from the second in Subsec. A.7. As mentioned above, because the fluctuation independence approximation is only usable in a midpoint-centered frame, $n(p,t)$ will refer always to this frame. $x(N,t)$ is well-defined without reference to any frame.

A.6.1 A number density master equation

If share-number fluctuations are independent at different $p$, a density $\pi(n,p,t)$ may be defined, which gives the probability to find $n$ orders in bin $(p,p+dp)$, at time $t$. The normalization condition defining $\pi$ as a probability density is

$$\sum_n \pi(n,p,t) = 1,$$  \hspace{1cm} (A.10)

for each bin index $p$ and at every $t$. The index $t$ will be suppressed henceforth in the notation since we are looking for time-independent solutions.

Supposing an arbitrary density of order-book configurations $\pi(n,p)$ at time $t$, the stochastic dynamics of the configurations causes probability to be redis-
tributed according to the master equation

\[ \frac{\partial}{\partial t} \pi(n,p) = \frac{\alpha(p) \, dp}{\sigma} \left[ \pi(n - \sigma,p) - \pi(n,p) \right] + \frac{\delta}{\sigma} \left[ (n + \sigma) \pi(n + \sigma,p) - n \pi(n,p) \right] + \frac{\mu(p)}{2\sigma} \left[ \pi(n + \sigma,p) - \pi(n,p) \right] + \sum_{\Delta p} P_+ (\Delta p) \left[ \pi(n,p - \Delta p) - \pi(n,p) \right] + \sum_{\Delta p} P_- (\Delta p) \left[ \pi(n,p + \Delta p) - \pi(n,p) \right]. \]  

(A.11)

Here \( \partial \pi(n,p)/\partial t \) is a continuum notation for \( [\pi(n,p,t + \delta t) - \pi(n,p,t)]/\delta t \), where \( \delta t \) is an elementary time step, chosen short enough that at most one event alters any typical configuration. Eq. (A.11) represents a general balance between additions and removals, without regard to the meaning of \( n \). Thus, \( \alpha(p) \) is a function that must be determined self-consistently with the choice of frame. As an example of how this works, in a bid-centered frame, \( \alpha(p) \) takes a fixed value \( \alpha(\infty) \) at all \( p \), because the deposition rate is independent of position and frame shifts. The midpoint-centered frame is more complicated, because depositions below the midpoint cause shifts that leave the deposited order above the midpoint. The specific consequence for \( \alpha(p) \) in this case will be considered below. \( \mu(p)/2 \) is, similarly, the rate of market orders surviving to cancel limit orders at price \( p \). \( \mu(p)/2 \) decreases from \( \mu(0)/2 \) at the ask (for buy market orders, because \( \mu \) total orders are divided evenly between buys and sells) to zero as \( p \to \infty \), as market orders are screened probabilistically by intervening limit orders. \( \alpha(\infty) \) and \( \mu(0) \) are thus the parameters \( \alpha \) and \( \mu \) of the simulation.

The lines of Eq. (A.11) correspond to the following events. The term proportional to \( \alpha(p) \, dp/\sigma \) describes depositions of discrete orders at that rate (because \( \alpha \) is expressed in \textit{shares per price per time}), which raise configurations from \( n - \sigma \) to \( n \) shares at price \( p \). The term proportional to \( \delta \) comes from deletions and has the opposite effect, and is proportional to \( n/\sigma \), the number of \textit{orders} that can independently decay. The term proportional to \( \mu(p)/2\sigma \) describes market order annihilations. For general configurations, the preceding three effects may lead to shifts of the origin by arbitrary intervals \( \Delta p \), and \( P_\pm \) are for the moment unknown distributions over the frequency of those shifts. They must be determined self-consistently with the configuration of the book which emerges from any solution to Eq. (A.11).

A limitation of the simple product representation of frame shifts is that it assumes that whole order-book configurations are transported under \( p \pm \Delta p \to p \), independently of the value of \( n(p) \). As long as fluctuations are independent, this is a good approximation for orders at all \( p \) which are not either the bid or the ask, either before or after the event that causes the shift. The correlations are never ignorable for the bins which are the bid and ask, though, and there
is some distribution of instances in which any \( p \) of interest plays those parts. Approximate methods to incorporate those correlations will require replacing the product form with a sum of products conditioned on states of the order book, as will be derived below.

The important point is that the order-flow dependence of Eq. (A.11) is independent of these self-consistency requirements, and may be solved by use of generating functionals at general \( \alpha(p) \), \( \mu(p) \), and \( P_{\pm} \). The solution, exact but not analytically tractable at general \( dp \), will be derived in closed form in the next subsection. It has a well-behaved continuum limit at \( dp \to 0 \), however, which is analytically tractable, so that special case will be considered in the following subsection.

### A.6.2 Solution by generating functional

The moment generating functional for \( \pi \) is defined for a parameter \( \lambda \in [0, 1] \), as

\[
\Pi(\lambda, p) \equiv \sum_{n/\sigma=0}^{\infty} \lambda^{n/\sigma} \pi(n, p) .
\]  

(A.12)

Introducing a shorthand for its value at \( \lambda = 0 \),

\[
\Pi(0, p) = \pi(0, p) \equiv \pi_0(p) ,
\]  

(A.13)

while the normalization condition (A.10) for probabilities gives

\[
\Pi(1, p) = 1 , \forall p .
\]  

(A.14)

By definition of the average of \( n(p) \) in the distribution \( \pi \), denoted \( \langle n(p) \rangle \),

\[
\frac{\partial}{\partial \lambda} \Pi(\lambda, p) \bigg|_{\lambda=1} = \frac{\langle n(p) \rangle}{\sigma},
\]  

(A.15)

and because \( \Pi \) will be regular in some sufficiently small neighborhood of \( \lambda = 1 \), one can expand

\[
\Pi(\lambda, p) = 1 + (\lambda - 1) \frac{\langle n(p) \rangle}{\sigma} + O(\lambda - 1)^2 .
\]  

(A.16)

Multiplying Eq. (A.11) by \( \lambda^{n/\sigma} \) and summing over \( n \), (and suppressing the argument \( p \) in the notation everywhere; \( \alpha(p) \) or \( \alpha(0) \) will be used where the distinction of the function from its boundary value is needed) the stationary solution for \( \Pi \) must satisfy

\[
0 = \frac{\lambda - 1}{\sigma} \left\{ \alpha d\Pi - \delta \sigma \frac{\partial \Pi}{\partial \lambda} - \frac{\mu}{2\lambda} (\Pi - \pi_0) \right\} \bigg|_{(\lambda, p)}
\]

\[
+ \sum_{\Delta p} P_+ (\Delta p) [\Pi(\lambda, p - \Delta p) - \Pi(\lambda, p)]
\]

\[
+ \sum_{\Delta p} P_- (\Delta p) [\Pi(\lambda, p + \Delta p) - \Pi(\lambda, p)] .
\]  

(A.17)
Only the symmetric case with no net drift will be considered here for simplicity, which requires $P_+ (\Delta p) = P_- (\Delta p) \equiv P (\Delta p)$. In a Fokker-Planck expansion, the (unrenormalized) diffusivity of whatever reference price is used as coordinate origin, is related to the distribution $P$ by

$$D \equiv \sum_{\Delta p} P (\Delta p) \Delta p^2. \quad (A.18)$$

The rate at which shift events happen is

$$R \equiv \sum_{\Delta p} P (\Delta p), \quad (A.19)$$

and the mean shift amount appearing at linear order in derivatives (relevant at $p \to 0$), is

$$\langle \Delta p \rangle \equiv \frac{\sum_{\Delta p} P (\Delta p) \Delta p}{\sum_{\Delta p} P (\Delta p)}. \quad (A.20)$$

Anywhere in the interior of the price range (where $p$ is not at any stage the bid, ask, or a point in the spread), Eq. (A.17) may be written

$$\left\{ \frac{\partial}{\partial \lambda} - \frac{D}{\delta (\lambda - 1)} \frac{\partial^2}{\partial p^2} - \frac{\alpha dp - \mu/2\lambda}{\delta \sigma} \right\} \Pi = \frac{\mu}{2\delta \sigma \lambda} \pi_0. \quad (A.21)$$

Evaluated at $\lambda \to 1$, with the use of the expansion (A.16), this becomes

$$\left( 1 - \frac{D}{\delta} \frac{\partial}{\partial p^2} \right) \langle n \rangle = \frac{\alpha dp}{\delta} - \frac{\mu}{2\delta} (1 - \pi_0). \quad (A.22)$$

At this point it is convenient to specialize to the case $dp \to 0$, wherein the eligible values of any $\langle n (p) \rangle$ become just $\sigma$ and zero. The expectation is then related to the probability of zero occupancy (at each $p$) as

$$\langle n \rangle = \sigma [1 - \pi_0], \quad (A.23)$$

yielding immediately

$$\frac{\alpha dp}{\delta} = \left[ \frac{\mu}{2\delta \sigma} + \left( 1 - \frac{D}{\delta} \frac{d^2}{dp^2} \right) \langle n \rangle \right]. \quad (A.24)$$

Eq. (A.24) defines the general solution $\langle n (p) \rangle$ for the master equation (A.11), in the continuum limit $2\alpha dp/\mu \to 0$. The shift distribution $P (\Delta p)$ appears only through the diffusivity $D$, which must be solved self-consistently, along with the otherwise arbitrary functions $\alpha$ and $\mu$. For the more general solution at large $dp$ refer to the appendix of [29].

A first step toward nondimensionalization may be taken by writing Eq. (A.24) in the form (re-introducing the indexing of the functions)

$$\frac{\alpha (p)}{\alpha (\infty)} = \left[ \frac{\mu (p)}{\mu (0)} + \epsilon \left( 1 - \frac{D}{\delta} \frac{d^2}{dp^2} \right) \frac{1}{\epsilon} \delta \langle n \rangle \right]. \quad (A.25)$$
Far from the midpoint, where only depositions and cancellations take place, orders in bins of width \( dp \) are Poisson distributed with mean \( \alpha (\infty) \frac{dp}{\delta} \). Thus, the asymptotic value of \( \delta \langle n \rangle / \alpha (\infty) dp \) at large \( p \) is unity. This is consistent with a limit for \( \alpha (p) / \alpha (\infty) \) of unity, and a limit for the screened \( \mu (p) / \mu (0) \) of zero. The reason for grouping the nondimensionalized number density with \( 1/\epsilon \), together with the proper normalization of the characteristic price scale, will come from examining the decay of the dimensionless function \( \mu (p) / \mu (0) \).

### A.6.3 Screening of the market-order rate

In the context of independent fluctuations, Eq. (A.23) implies a relation between the mean density and the rate at which market orders are screened as price increases. The effect of a limit order, resident in the price bin \( p \) when a market order survives to reach that bin, is to prevent its arriving at the bin at \( p + dp \). Though the nature of the shift induced, when such annihilation occurs, depends on the comoving frame being modeled, the change in the number of orders surviving is independent of frame, and is given by

\[
d\mu = -\mu (1 - \pi_0) = -\mu \langle n \rangle / \sigma.
\]

Eq. (A.26) may be rewritten

\[
\frac{d \log (\mu (p) / \mu (0))}{dp} = -\frac{1}{\epsilon} \left( \frac{2\alpha (\infty)}{\mu (0)} \right) \left( \frac{\delta \langle n (p) \rangle}{\alpha (\infty) dp} \right),
\]

identifying the characteristic scale for prices as \( p_c = \mu (0) / 2\alpha (\infty) \equiv \mu / 2\alpha \). Writing \( \hat{p} \equiv p/p_c \), the function that screens market orders is the same as the argument of Eq. (A.25), and will be denoted

\[
\frac{1}{\epsilon} \frac{\delta \langle n (p) \rangle}{p_{\alpha (\infty)} dp} \equiv \psi (\hat{p})
\]

Defining a nondimensionalized diffusivity \( \beta \equiv D / \delta p_c^2 \), Eq. (A.24) can then be put in the form

\[
\frac{\alpha (p)}{\alpha (\infty)} = \left[ \frac{\mu (p)}{\mu (0)} + \epsilon \left( 1 - \beta \frac{d^2}{dp^2} \right) \right] \psi,
\]

with

\[
\frac{\mu (p)}{\mu (0)} \equiv \varphi (\hat{p}) = \exp \left( -\int_0^{\hat{p}} d\hat{p}' \psi (\hat{p}') \right),
\]

### A.6.4 Verifying the conservation laws

Since nothing about the derivation so far has made explicit use of the frame in which \( n \) is averaged, the combination of Eq. (A.29) with Eq. (A.30) respects the conservation laws (A.6) and (A.7), if appropriate forms are chosen for the deposition rate \( \alpha (p) \).
For example, in the bid-centered frame, \( \alpha(p)/\alpha(\infty) = 1 \) everywhere. Multiplying Eq. (A.29) by \( d\hat{p} \) and integrating over the whole range from the bid to \(+\infty\), we recover the nondimensionalized form of Eq. (A.6):
\[
\int_0^\infty d\hat{p} (1 - \epsilon \psi) = 1, \quad (A.31)
\]
if we are careful with one convention. The integral of the diffusion term formally produces the first derivative \( d\psi/d\hat{p} \). We must regard this as a true first derivative, and consider its evaluation at zero continued far enough below the bid to capture the identically zero first derivative of the sell order depth profile.

In the midpoint centered frame, the correct form for the source term should be \( \alpha(\hat{p})/\alpha(\hat{\infty}) = 1 + \text{Pr}(\hat{s}/2 \geq \hat{p}) \), whatever the expression for the cumulative distribution function. Recognizing that the integral of the CDF is, by parts, the mean value of \( \hat{s}/2 \), the same integration of Eq. (A.29) gives
\[
\int_0^\infty d\hat{p} (1 - \epsilon \psi) = 1 - \langle \hat{s} \rangle/2, \quad (A.32)
\]
the nondimensionalized form of Eq. (A.7). Again, this works only if the surface contribution from integrating the diffusion term vanishes.

Neither of these results required the assumption of independent fluctuations, though that will be used below to give a simple approximate form for \( \text{Pr}(\hat{s}/2 \geq \hat{p}) \approx \varphi(\hat{p}) \). They therefore, provide a check that the extinction form (A.30) propagates market orders correctly into the interior of the order-book distribution, to respect global conservation. They also check the consistency of the intuitively plausible form for \( \alpha \) in the midpoint-centered frame. The detailed form is then justified whenever the assumption of independent fluctuations is checked to be valid.

### A.6.5 Self-consistent parametrization

The assumption of independent fluctuations of \( n(p) \) used above to derive the screening of market orders, is equivalent to a specification of the CDF of the ask. Market orders are only removed between prices \( p \) and \( p + dp \) in those instances when the ask is at \( p \). Therefore,
\[
\text{Pr}(\hat{s}/2 \geq \hat{p}) = \varphi(\hat{p}), \quad (A.33)
\]
the continuum limit of Eq. (A.9). Together with the form \( \alpha(\hat{p})/\alpha(\infty) = 1 + \text{Pr}(\hat{s}/2 \geq \hat{p}) \), Eq. (A.29) becomes
\[
1 + \varphi = -\left[ \frac{d\varphi}{d\hat{p}} + \epsilon \left( 1 - \beta \frac{d^2}{d\hat{p}^2} \right) \frac{d\log \varphi}{d\hat{p}} \right]. \quad (A.34)
\]
(If the assumption of independent fluctuations were valid in the bid-centered frame, it would take the same form, but with \( \varphi \) removed on the left-hand side.)

To consistently use the diffusion approximation, with the realization that for \( p = 0, n\pi(n, p - \Delta p) = 0 \) for essentially all \( \Delta p \) in Eq. (A.11), it is necessary to set
the Fokker-Planck approximation to \( \psi(0 - \langle \Delta p \rangle) = 0 \) as a boundary condition. Nondimensionalized, this gives
\[
\frac{\beta}{2} \frac{d^2 \psi}{d \hat{p}^2} \bigg|_0 = \frac{R}{\delta} \left( \frac{\langle \Delta \hat{p} \rangle}{d \hat{p}} - 1 \right) \psi \bigg|_0,
\]  
(A.35)
where \( R \) is the rate at which shifts occur (Eq. A.19). In the solutions below, the curvature will typically be much smaller than \( \psi(0) \sim 1 \), so it will be convenient to enforce the simpler condition
\[
\langle \Delta \hat{p} \rangle \frac{d \psi}{d \hat{p}} \bigg|_0 - \psi(0) \approx 0,
\]  
(A.36)
and verify that it is consistent once solutions have been evaluated.

Self-consistent expressions for \( \beta \) and \( \langle \Delta \hat{p} \rangle \) are then constructed as follows. Given an ask at some position \( a \) (in the midpoint-centered frame), there is a range from \(-a\) to \( a\) in which sell limit orders may be placed, which will induce positive midpoint-shifts. The shift amount is half as great as the distance from the bid, so the measure for shifts \( dP_+ (\Delta \hat{p}) \) from sell limit-order addition inherits a term \( 2\alpha(0) (d \Delta p) \Pr(a \geq \Delta p) \), where the last factor counts the instances with asks large enough to admit shifts by \( \Delta p \). There is an equal contribution to \( dP_- \) from the additions of buy limit orders. Symmetry requires that for every positive shift due to an addition, there is a negative shift due to evaporation with equal measure, so the contribution from buy limit order removal should equal that for sell limit order addition. When these contributions are summed, the measures for positive and negative shifts both equal
\[
dP_{\pm} (\Delta \hat{p}) = 4\alpha(\infty) (d \Delta p) \Pr(a \geq \Delta p).
\]  
(A.37)

Eq. (A.37) may be inserted into the continuum limit of the definition (A.18) for \( D \), and then nondimensionalized to give
\[
\beta = \frac{4}{\epsilon} \int_0^\infty d\Delta \hat{p} (\Delta \hat{p})^2 \varphi(\Delta \hat{p}),
\]  
(A.38)
where the mean-field substitution of \( \varphi(\Delta \hat{p}) \) for \( \Pr(a \geq \Delta p) \) has been used. Similarly, the mean shift amount used in Eq. (A.36) is
\[
\langle \Delta \hat{p} \rangle = \frac{\int_0^\infty d\Delta \hat{p} (\Delta \hat{p}) \varphi(\Delta \hat{p})}{\int_0^\infty d\Delta \hat{p} \varphi(\Delta \hat{p})}.
\]  
(A.39)

A fit of Eq. (A.34) to simulations, using these self-consistent measures for shifts, is shown in Figure A.5. This solution is actually a compromise between approximations with opposing ranges of validity. The diffusion equation using the mean order depth describes the nonzero transport of limit orders through the midpoint, an approximation inconsistent with the correlations of shifts with states of the order book. This approximation is a small error only at \( \epsilon \to 0 \). On the other hand, both the form of \( \alpha \), and the self-consistent solutions for \( \langle \Delta \hat{p} \rangle \) and \( \beta \), made use of the mean-field approximation, which we saw was only valid for \( \epsilon \ll 1 \). The two approximations appear to create roughly compensating errors in the intermediate range \( \epsilon \sim 0.02 \).
A.6.6 Accounting for correlations

The numerical integral implementing the diffusion solution actually doesn’t satisfy the global conservation condition that the diffusion term integrate to zero over the whole price range. Thus, it describes diffusive transport of orders through the midpoint, and as such also doesn’t have the right $\hat{p} = 0$ boundary condition. The effective absorbing boundary represented by the pure diffusion solution corresponds roughly to the approximation made by Bouchaud et al. [50] It differs from theirs, though, in that their method of images effectively approximates the region of the spread as a point, whereas Eq. (A.29) actually resolves the screening of market orders as the spread fluctuates.

Treating the spread region – roughly defined as the range over which market orders are screened – as a point is consistent with treating the resulting coarse-grained “midpoint” as an absorbing boundary. If the spread is resolved, however, it is not consistent for diffusion to transport any finite number density through the midpoint, because the midpoint is always strictly in the center of an open set with no orders, in a continuous price space. The correct behavior in a neighborhood of the “fine-grained midpoint” can be obtained by explicitly accounting for the correlation of the state of orders, with the shifts that are produced when market or limit order additions occur.

We expect the problem of recovering both the global conservation law and the correct $\hat{p} = 0$ boundary condition to be difficult, as it should be responsible for the non-trivial corrections to short-term and long-term diffusion mentioned earlier. We have found, however, that by explicitly sacrificing the global conservation law, we can incorporate the dependence of shifts on the position of the ask, in an interesting range around the midpoint. At general $\epsilon$, the corrections to diffusion reproduce the mean density over the main support of the CDF of the spread. While the resulting density does not predict that CDF (due to correlated fluctuations), it closely enough resembles the real density that the independent CDFs of the two are similar.

A.6.7 Generalizing the shift-induced source terms

Nondimensionalizing the generating-functional master equation (A.17) and keeping leading terms in $dp$ at $\lambda \to 1$, get

$$\frac{\alpha (\hat{p})}{\alpha (\infty)} = \left( \frac{\mu (\hat{p})}{\mu (0)} + \epsilon \right) \psi (\hat{p})$$

$$\quad - \int dP_+ (\Delta \hat{p}) [\psi (\hat{p} - \Delta \hat{p}) - \psi (\hat{p})]$$

$$\quad - \int dP_- (\Delta \hat{p}) [\psi (\hat{p} + \Delta \hat{p}) - \psi (\hat{p})]$$

(A.40)

where $dP_\pm (\Delta \hat{p})$ is the nondimensionalized measure that results from taking the continuum limit of $P_\pm$ in the variable $\Delta \hat{p}$.

Eq. (A.40) is inaccurate because the number of orders shifted into or out of a price bin $p$, at a given spread, may be identically zero, rather than the
unconditional mean value $\psi$. We take that into account by replacing the last two lines of Eq. (A.40) with lists of source terms, whose forms depend on the position of the ask, weighted by the probability density for that ask. Independent fluctuations are assumed by using Eq. (A.33).

It is convenient at this point to denote the replacement of the last two lines of Eq. (A.40) with the notation $S$, yielding

$$\frac{\alpha(\hat{p})}{\alpha(\infty)} = \left(\frac{\mu(\hat{p})}{\mu(0)} + \epsilon\right) \psi - S.$$  \hspace{1cm} (A.41)

The global conservation laws for orders would be satisfied if $\int d\hat{p} S = 0$.

The source term $S$ can be derived approximately. (See the appendix of Smith et al. [29].) The solution to Eq. (A.41) at $\epsilon = 0.2$ is compared to the simulated order-book depth and spread distribution in Figure A.6. The simulated $\langle n(p) \rangle$ satisfies Eq. (A.32), showing the correct "remainder area" below the line $\langle n \rangle \equiv 1$. The numerical integral deviates from that value by the incorrect integral $\int d\hat{p} S \neq 0$. However, most of the probability for the spread lies within the range where the source terms $S$ are approximately correct, and as a result the distribution for $\hat{s}/2$ is predicted fairly well.

Even where the mean-field approximation is known to be inadequate, the source terms defined here capture most of the behavior of the order-book distribution in the region that affects the spread distribution. Figure A.7 shows the comparison to simulations for $\epsilon = 0.02$, and Figure A.8 for $\epsilon = 0.002$. Both cases fail to reproduce the distribution of the spread, and also fail to capture the large-$\hat{p}$ behavior of $\psi$. However, they approximate $\psi$ at small $\hat{p}$ well enough that the resulting distribution of the spread is close to what would be produced by the simulated $\psi$ if fluctuations were independent.

### A.7 A mean-field theory of order separation intervals: the Independent Interval Approximation

A simplifying assumption that is in some sense dual to independent fluctuations of $n(p)$, is independent fluctuations in the intervals $x(N)$ at different $N$. Here we develop a mean-field theory for the order separation intervals in this model. From this, we will also be able to make an estimate of the depth profiles for any value of the parameters. For convenience of notation we will use $x_N$ to denote $x(N)$.

Limit order placements are considered to take place strictly on sites which are not occupied. This is the same level of approximation as made in the previous section. The time step is normalized to unity, as above, so that rates are equal to probabilities after one update of the whole configuration. The rates $\alpha$ and $\mu$ used in this section correspond to $\alpha(\infty)$ and $\mu(0)$ as defined earlier.

As shown in Figure A.1 the configuration is entirely specified instant by instant if the instantaneous values of the order separation intervals are known.
Consider now, how these intervals might change due to various processes. For the spread $x_0$, these processes and the corresponding change in $x_0$, are listed below.

1. $x_0 \rightarrow x_0 + x_1$ with rate $(\delta + \mu/2\sigma)$ (when the ask either evaporates or is deleted by a market order).

2. $x_0 \rightarrow x_0 + x_{-1}$ with rate $(\delta + \mu/2\sigma)$ (when the bid either evaporates or is deleted by the corresponding market order).

3. $x_0 \rightarrow x'$ for any value $1 \leq x' \leq x_0 - 1$, when a sell limit order is deposited anywhere in the spread. The rate for any single deposition is $\alpha dp/\sigma$, so the cumulative rate for some deposition is $\alpha dp (x_0 - 1)/\sigma$. (The $-1$ comes from the prohibition against depositing on occupied sites.)

4. Similarly $x_0 \rightarrow x_0 - x'$ for any $1 \leq x' \leq x_0 - 1$, when a buy limit order is deposited in the spread, also with cumulative rate $\alpha dp (x_0 - 1)/\sigma$.

5. Since the above processes describe all possible single-event changes to the configuration, the probability that it remains unchanged in a single time step is $1 - 2\delta - \mu/\sigma - 2\alpha dp (x_0 - 1)/\sigma$.

In all that follows, we will put $\sigma = 1$ without loss of generality. If we know $x_0$, $x_1$, and $x_{-1}$ at time $t$, the expected value at time $t + dt$ is then

$$x_0(t + dt) = x_0(t) [1 - 2\delta - \mu_0 - 2a (x_0 - 1)] + (x_0 + x_1) \left( \delta + \frac{\mu}{2} \right) + (x_0 + x_{-1}) \left( \delta + \frac{\mu}{2} \right) + (\alpha dp)x_0(x_0 - 1)$$

(A.42)

Here, $x_i(t)$ represents the value of the interval averaged over many realizations of the process evolved up to time $t$. Again representing the finite difference as a time derivative, the change in the expected value, given $x_0$, $x_1$, and $x_{-1}$, is

$$\frac{dx_0}{dt} = (x_1 + x_{-1}) \left( \delta + \frac{\mu}{2} \right) - (\alpha dp)x_0(x_0 - 1).$$

(A.43)

Were it not for the quadratic term arising from deposition, Eq. (A.43) would be a linear function of $x_0$, $x_1$, and $x_{-1}$. However, we now need an approximation for $\langle x_0^2 \rangle$, where the angle brackets represent an average over realizations as before or equivalently a time average in the steady state. Let us for the moment assume that we can approximate $\langle x_0^2 \rangle$ by $a \langle x_0 \rangle^2$, where $a$ is some as yet undetermined constant to be determined self-consistently. We will make this approximation for all the $x_k$’s. This is clearly not entirely accurate because the PDF of $x_k$ could depend on $k$. (Indeed, it does and we will comment on this a little later.) However, as we will see, this is still a very good approximation.

We will, therefore, make this approximation in Eq. (A.43) and everywhere below, and look for steady state solutions when the $x_k$’s have reached a time independent average value.
case | rate | range  \\
---|---|---  \\
x_1 \rightarrow x_2 | \delta + \mu/2 | x' \in (1, x_0 - 1)  \\
x_1 \rightarrow (x_1 + x_2) | \delta | x' \in (1, x_0 - 1)  \\
x_1 \rightarrow x' | odp | x' \in (1, x_1 - 1)  \\
x_1 \rightarrow x_1 - x' | odp | x' \in (1, x_1 - 1)  \\

Table A.1: Events that can change the value of \(x_1\), with their rates of occurrence.

It then follows that,

\[
(\delta + \mu/2) (x_1 + x_{-1}) = aodpx_0 (x_0 - 1) \quad (A.44)
\]

The interval \(x_k\) may be thought of as the inverse of the density at a distance \(\sum_{j=0}^{k-1} x_j\) from the bid. That is, \(x_i \approx 1/\left<n \left(\sum_{j=0}^{i-1} x_j dp\right)\right>\), the dual to the mean depth, at least at large \(i\). It therefore, makes sense to introduce a normalized interval

\[
\hat{x}_i \equiv \frac{\epsilon}{\delta} x_i dp = \frac{x_i dp}{p_c} \approx \frac{1}{\psi \left(\sum_{j=0}^{i-1} \hat{x}_j\right)}, \quad (A.45)
\]

the mean-field inverse of the normalized depth \(\psi\). In this nondimensionalized form, Eq. \((A.44)\) becomes

\[
(1 + \epsilon) (\hat{x}_1 + \hat{x}_{-1}) = a\hat{x}_0 (\hat{x}_0 - d\hat{p}), \quad (A.46)
\]

where \(d\hat{p} = dp/p_c\).

Since the depth profile is symmetric about the origin, \(\hat{x}_1 = \hat{x}_{-1}\). From the equations, it can be seen that this ansatz is self-consistent and extends to all higher \(\hat{x}_i\). Substituting this in Eq. \((A.46)\) we get

\[
(1 + \epsilon) \hat{x}_1 = \frac{a}{2} \hat{x}_0 (\hat{x}_0 - d\hat{p}) = (1 + \epsilon) \hat{x}_{-1}. \quad (A.47)
\]

Proceeding to the change of \(x_1\), the events that can occur, with their probabilities, are shown in Table A.1, with the remaining probability that \(x_1\) remains unchanged.

The differential equation for the mean change of \(x_1\) can be derived along previous lines and becomes

\[
\frac{dx_1}{dt} = \left(2\delta + \frac{\mu}{2}\right) x_2 - \left(\delta + \frac{\mu}{2}\right) x_1 + odp \left[\frac{x_0 (x_0 - 1)}{2} - \frac{x_1 (x_1 - 1)}{2} - x_1 (x_0 - 1)\right]. \quad (A.48)
\]

Note that in the above equations, the mean-field approximation consists of assuming that terms like \(\langle x_0 x_1 \rangle\) are approximated by the product \(\langle x_0 \rangle \langle x_1 \rangle\). This is thus an ‘independent interval’ approximation.
Nondimensionalizing Eq. (A.48) and combining the result with Eq. (A.47) gives the stationary value for $x_2$ from $x_0$ and $x_1$,

$$(1 + 2\epsilon) \hat{x}_2 = \frac{a}{2} \hat{x}_1 (\hat{x}_1 - d\hat{p}) + \hat{x}_1 (\hat{x}_0 - d\hat{p}).$$

(A.49)

Following the same procedure for general $k$, the nondimensionalized recursion relation is

$$(1 + k\epsilon) \hat{x}_k = \frac{a}{2} \hat{x}_{k-1} (\hat{x}_{k-1} - d\hat{p}) + \hat{x}_{k-1} \sum_{i=0}^{k-2} (\hat{x}_i - d\hat{p}).$$

(A.50)

### A.7.1 Asymptotes and conservation rules

Far from the bid or ask, $\hat{x}_k$ must go to a constant value, which we denote $\hat{x}_\infty$. In other words, for large $k$, $\hat{x}_{k+1} \rightarrow \hat{x}_k$. Taking the difference of Eq. (A.50) for $k+1$ and $k$ in this limit gives the identification

$$\epsilon \hat{x}_\infty = \hat{x}_\infty (\hat{x}_\infty - d\hat{p}),$$

(A.51)

or $\hat{x}_\infty = \epsilon + d\hat{p}$. Apart from the factor of $d\hat{p}$, arising from the exclusion of deposition on already-occupied sites, this agrees with the limit $\psi(\infty) \rightarrow 1/\epsilon$ found earlier. In the continuum limit $d\hat{p} \rightarrow 0$ at fixed $\epsilon$, these are the same.

From the large-$k$ limit of Eq. (A.50), one can also solve easily for the quantity $S_\infty \equiv \sum_{i=0}^{\infty} (\hat{x}_i - \hat{x}_\infty)$, which is related to the bid-centered order conservation law mentioned in Section A.3. Dividing by a factor of $\hat{x}_\infty$ at large $k$,

$$1 + (1 - a/2)\epsilon = S_\infty.$$

(A.53)

The interpretation of $S_\infty$ is straightforward. There are $k+1$ orders in the price range $\sum_{i=0}^{k} x_i$. Their decay rate is $\delta (k+1)$, and the rate of annihilation from market orders is $\mu/2$. The rate of additions, up to an uncertainty about what should be considered the center of the interval, is $(\alpha d\hat{p}) \sum_{i=0}^{k} (x_i - 1)$ in the bid-centered frame (where effective $\alpha$ is constant and additions on top of previously occupied sites is forbidden). Equality of addition and removal is the bid-centered order conservation law (again), in the form

$$\frac{\mu}{2} + \delta (k+1) = \alpha d\hat{p} \sum_{i=0}^{k} (x_i - 1).$$

(A.54)

Taking $k$ large, nondimensionalizing, and using Eq. (A.51), Eq. (A.54) becomes

$$(1 + a/2)\epsilon = S_\infty.$$

(A.55)
This conservation law is indeed respected to a remarkable accuracy in Monte Carlo simulations of the model as indicated in Table A.2.

In order that the equation for the $x$’s obey this exact conservation law, we require Eq. A.53 to be equal to Eq. A.55. We can hence now self-consistently set the value of $a = 2$.

The value of $a$ implies that we have now set $\langle x_k^2 \rangle \sim 2 \langle x_k \rangle^2$. This would be strictly true if the probability distribution function of the interval $x_k$ were exponentially distributed for all $k$. This is generally a good approximation for large $k$ for any $\epsilon$. Figure A.9 shows the numerical results from Monte Carlo simulations of the model, for the probability distribution function for three intervals $x_0$, $x_1$ and $x_5$ at $\epsilon = 0.1$. The functional form for $P(x_0)$ and $P(x_1)$ are better approximated by a Gaussian than an exponential. However, $P(x_5)$ is clearly an exponential.

Eq. (A.55) has an important consequence for the short-term and long-term diffusivities, which can also be seen in simulations, as mentioned in earlier sections. The nondimensionalization of the diffusivity $D$ with the rate parameters, suggests a classical scaling of the diffusivity

$$D \sim p_c^2 \delta = \frac{\mu^2}{4\alpha^2}\delta.$$  \hspace{1cm} (A.56)

As mentioned earlier, it is observed from simulations that the locally best short-time fit to the actual diffusivity of the midpoint is $\sim \sqrt{1/\epsilon}$ times the estimate (A.56), and the long-time diffusivity is $\sim \sqrt{\epsilon}$ times the classical estimate. While we do not yet know how to derive this relation analytically, the fact that early and late-time renormalizations must have this qualitative relation can be argued from the conservation law (A.55).

$S_\infty$ is the area enclosed between the actual density and the asymptotic value. Increases in $1/\epsilon$ (descaled market-order rate) deplete orders near the spread, diminishing the mean depth at small $\hat{p}$, and induce the upward curvature seen in Figure 2.3, and even more strongly in Figure A.14 below. As noted above, they cause more frequent shifts (more than compensating for the slight decrease in average step size), and increase the classically descaled diffusivity $\beta$. However, as a result, this increases the fraction of the area in $S_\infty$ accumulated near the spread, requiring that the mean depth at larger $\hat{p}$ increase to compensate (see Figure A.14). The resulting steeper approach to the asymptotic depth at prices greater than the mean spread, and the larger negative curvature of the distribution, are fit by an effective diffusivity that decreases with increasing $1/\epsilon$. Since

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$S_\infty$ from theory</th>
<th>$S_\infty$ from MCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.66</td>
<td>1</td>
<td>1.000</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>1.000</td>
</tr>
<tr>
<td>0.04</td>
<td>1</td>
<td>0.998</td>
</tr>
<tr>
<td>0.02</td>
<td>1</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table A.2: Theoretical vs. results from simulations for $S_\infty$. 
the distribution further from the midpoint represents the imprint of market order activity further in the past, this effective diffusivity describes the long-term evolution of the distribution. The resulting anticorrelation of the small-\( \hat{p} \) and large-\( \hat{p} \) effective diffusion constants implied by conservation of the area \( S_\infty \) is exactly consistent with their respective \( \sim \sqrt{1/\epsilon} \) and \( \sim \sqrt{\epsilon} \) scalings. The general idea here is to connect diffusivities at short and long time scales to the depth profile near the spread and far away from the spread respectively. The conservation law for the depth profile then implies a connection between these two diffusivities.

A.7.2 Direct simulation in interval coordinates

The set of equations determined by the general form (A.50) is ultimately parametrized by the single input \( \hat{x}_0 \). The correct value for \( \hat{x}_0 \) is determined when the \( \hat{x}_k \) are solved recursively, by requiring convergence to \( \hat{x}_\infty \). We do this recursion numerically, in the same manner as was done to solve the differential equation for the normalized mean density \( \psi(\hat{p}) \).

In Figure A.10 we compare the numerical result for \( \hat{x}_0 \) with the analytical estimate generated as explained above. The results are surprisingly good throughout the entire range. Though the theoretical value consistently underestimates the numerical value, yet the functional form is captured accurately.

In Figure A.11, the values of \( x_k \) for all \( k \), are compared to the values determined directly from simulations.

Figure A.12 shows the same data on a semilog scale for \( \hat{x}_k/\hat{x}_\infty - 1 \), showing the exponential decay at large argument characteristic of a simple diffusion solution. The IIA is clearly a good approximation for large \( \epsilon \). However, for small \( \epsilon \) it starts deviating significantly from the simulations, especially for large \( k \).

The values of \( x_k \) computed from the IIA, can be very directly used to get an estimation of the price impact. The price impact, as defined in earlier sections, can be thought of as the change in the position of the midpoint (or the bid), consecutive to a certain number of orders being filled. Within the framework of the simplified model we study here, this is simply the quantity \( \langle \Delta m \rangle = 1/2 \sum_{k'=1}^{k} x_{k'} \), for \( k \) orders. The factor of 1/2 comes from considering the change in the position of the midpoint and not the bid. Figure A.13 shows \( \langle \Delta m \rangle \) nondimensionalized by \( p_c \) plotted as a function of the number of orders (multiplied by \( \epsilon \)), for three different values of \( \epsilon \). Again, the theory matches quite well with the numerics, qualitatively. For large \( \epsilon \) the agreement is quantitative as well.

The simplest approximation to the density profiles in the midpoint-centered frame is to continue to approximate the mean density as \( 1/x_k \), but to regard that density as evaluated at position \( x_0/2 + \sum_{k=1}^{k} x_k \). This clearly is not an adequate treatment in the range of the spread, both because the intervals are discrete, whereas mean \( \psi \) is continuous, and because the density profiles satisfy different global conservation laws associated with non-constancy of \( \alpha \). For large \( k \) however, this approximation might hold. The mean-field values (only) corresponding to a plot of \( \epsilon\psi(\hat{p}) \) versus \( \hat{p} \), are shown in Figure A.14. Here the theoretically estimated \( x_k \)'s at different parameter values are used to generate the depth profile using the procedure detailed above.
A comparison of the theoretically estimated profiles with the results from Monte Carlo simulations of the model, is shown in Figure A.15. As evident, the theoretical estimate for the density profile is better for large \( \epsilon \) than for small \( \epsilon \).

We can also generalize the above analysis to when the order placement process is no longer uniform. In particular it has been found that a power-law order placement process is relevant [51, 50]. We carry out the above analysis for when \( \alpha = \Delta_0^\beta / (\Delta + \Delta_0)^\beta \) where \( \Delta \) is the distance from the current bid and \( \Delta_0 \) determines the ‘shoulder’ of the power-law. We find an interesting dependence of the existence of solutions on \( \beta \). In particular we find that for \( \beta > 1 \), \( \Delta_0 \) needs to be larger than some value (which depends on \( \beta \) as well as other parameters of the model such as \( \mu \) and \( \delta \)) for solutions of the IIA to exist. This might be interpreted as a market order wiping out the entire book, if the exponent is too large. When solutions exist, we find that the depth profile has a peak, consistent with the findings of Bouchaud et al. [50]. In Figure A.16 the depth profiles for three different values of \( \Delta_0 \) are plotted.
Figure A.4: CDFs $\Pr(s/2 < p)$ from simulations (thin solid), mean density profile $\langle n_m(p) \rangle$ from simulations (thick solid), and computed CDF of spread (thin dashed) from $\langle n_m(p) \rangle$, under the assumption of uncorrelated fluctuations, at three values of $\epsilon$: (a) $\epsilon = 0.2$ (low market order rate), approximation is very good; (b) $\epsilon = 0.02$ (intermediate market order rate), approximation is marginal; (c) $\epsilon = 0.002$ (high market order rate), approximation is very poor.
Figure A.5: Fit of the self-consistent solution with diffusivity term to simulation results for the midpoint-centered frame. Thin solid line is the analytic solution for the mean number density, and thick solid line is the simulation result, at $\epsilon = 0.02$. Thin dashed line is the analytic prediction for the cumulative distribution function $\text{Pr}(\hat{s}/2 \leq \hat{p})$, and thick dashed line is the simulation result.

Figure A.6: Reconstruction with source terms $S$ that approximately account for correlated fluctuations near the midpoint. $\epsilon = 0.2$. Thick solid line is averaged order book depth from simulations, and thin solid line is the mean field result. Thin dotted line is the simulated CDF for $\hat{s}/2$, and thick dotted line is the mean field result. Thick dashed line is the CDF that would be produced from the simulated depth, if the mean-field approximation were exact.
Figure A.7: Reconstruction with correlated source terms for $\epsilon = 0.02$. Line style and thickness are the same as in Figure A.6.

Figure A.8: Reconstruction with correlated source terms for $\epsilon = 0.002$. Line style and thickness are the same as in Figure A.6.
Figure A.9: The probability distribution functions $P_x(y)$ vs. $y$ for the intervals $x = x_0, x_1$ and $x_5$ at $\epsilon = 0.1$, on a semi-log scale. Solid curve is for $x_0$, dashed for $x_1$, and dot for $x_5$. The functional form of the distribution changes from a Gaussian to an exponential.

Figure A.10: The mean value of the spread in nondimensional units $\hat{s} = s/p_c$ as a function of $\epsilon$. The numerical value above (solid) is compared with the theoretical estimate below (dash).
Figure A.11: Four pairs of curves for the quantity $\hat{x}_k/\hat{x}_\infty - 1$ vs. $k$. The value of $\epsilon$ increases from top to bottom ($\epsilon = 0.02, 0.04, 0.2, 0.66$). In each pair of curves, the markers are obtained from simulations while the solid curve is the prediction of Eq. A.50 evaluated numerically. The difference between numerics and mean-field increases as $\epsilon$ decreases, especially for large $k$.

Figure A.12: Same plot as Figure A.11 but on a semi-log scale to show exponential decay at large $k$. 

122
Figure A.13: Three pairs of curves for the quantity $\langle \Delta m \rangle / p_c$ vs. $N\epsilon$ where $\langle \Delta m \rangle = 1/2 \sum_{k=1}^{N} x_k$. The value of $\epsilon$ increases from top to bottom ($\epsilon = 0.002, 0.02, 0.2$). In each pair of curves, the markers are obtained from simulations while the solid curve is the prediction of the IIA. For $\epsilon = 0.002$, we show only the theoretical prediction. The theory captures the functional form of the price impact curves for different $\epsilon$. Quantitatively, it is better for larger $\epsilon$, as remarked earlier.

Figure A.14: Density profiles for different values of $\epsilon$ ranging over the values 0.2, 0.02, 0.004, 0.001, obtained from the Independent Interval Approximation.
Figure A.15: Density profiles from Monte Carlo simulation (markers) and the Independent Interval Approximation (lines). Pluses and dash line are for $\epsilon = 0.2$, while crosses and dotted line are for $\epsilon = 0.02$.

Figure A.16: Density profiles for a power-law order placement process for different values of $\Delta_0$. 

124
Bibliography


[17] Budapest Stock Exchange LTD., Regulation of the Budapest Stock Exchange LTD. on the code of trading 


131


