Market Microstructure Theory and Strategic Behavior of Market Makers

Doctoral Thesis

Presented to the Faculty of Economic and Social Sciences at the University of Fribourg (Switzerland) to obtain the doctorate degree in Economic and Social Sciences

by

Andreas Krause

Accepted by the Faculty of Economic and Social Sciences on 12th October 2000 on the recommendation of

Professor Dr. Gerhard Aschinger (First Adviser)

and

Professor Dr. Christoph Kaserer (Second Adviser)

Bath, Great Britain 2000

"Among the plays which men perform in taking different parts in this magnificent world theater, the greatest comedy is played at the exchange. There, ... hiding places, concealment of facts, quarrels, provocations, mockery, idle talk, violent desires, collusion, artful deception, betrayals, cheatings, and even tragic end are to be found."

Joseph de la Vega, *Confusion de Confusiones* (1688)

Contents

List of Figures

List of Tables

List of Abbreviations

ACES	Advanced Computerized Execution System
AMEX	American Exchange
CATS	Computer Assisted Trading System
CBoT	Chicago Board of Trade
CHX	Chicago Stock Exchange
CME	Chicago Mercantile Exchange
DOT	Designated Order Turnaround
DTB	Deutsche Terminbörse
EBS	Elektronische Börse Schweiz
ECN	Electronic Communications Network
EUREX	European Exchange
FIPS	Fixed Income Pricing System
IBIS	Inter-Banken-Informations-System
IPO	Initial Public Offering
iX	International Exchanges
LIFFE	London International Financial Futures Exchange
LSE	London Stock Exchange
MATIF	Marché à Terme des Instruments Financiers de Paris

NASD	National Association of Securities Dealers
NASDAQ	National Association of Securities Dealers Automated Quotation System
NASDR	NASD Regulation, Inc.
NMS	National Market System
NODES	NASDAQ Order Delivery and Execution System
NYSE	New York Stock Exchange
OCT	Order Confirmation Transaction Service
OTC	Over-the-Counter
OTCBB	OTC Bulletin Board
PHLX	Philadelphia Stock Exchange
PORTAL	Private Offerings, Resales and Trading through Automated Linkages
SEC	Securities and Exchange Commission
SOES	Small Order Execution System
SOFFEX	Swiss Options and Financial Futures Exchange
TSE	Toronto Stock Exchange
USD	US-Dollar
VSE	Vienna Stock Exchange
XETRA	Exchange Electronic Trading

Chapter 1

Introduction

In recent years the attitude towards stocks has changed. Before, they were regarded mainly as a mean to invest for wealthy individuals, while others were restricted to bank and savings accounts. Therefore information on stock markets received only very limited attention. Nowadays news from stock markets are incorporated into the daily news on television or radio and a large number of newspapers and magazines are published, distributing information as well as investment advice. This development, observed throughout Northern America and Europe, gives evidence of the interest a large number of individuals have in stock markets.

We may identify two main reasons for this development. Firstly, in Europe as well as in the United States, since 1982 the stock market experienced a long upswing. By investing into stock markets one could get an annual return of approximately 15% when investing in stocks of the Dow Jones Industrial Average and even 20% p.a. on the NASDAQ, compared to interest rates of significant less than 10% p.a. With exception of the crash in 1987, stock markets experienced no longer periods of stagnation or even negative returns, unlike in the 1960s and 1970s. For this reason individuals were more and more willing to invest their wealth into the stock market in order to profit from these high returns. We observed this change to start in the United States in the 1980s, while Europe has been affected only in the middle of the 1990s. The predicted growth of companies with internet activities further accelerated this change since 1998.

A second cause for this development is more significant in Europe than in the United States. Severe problems with social security, especially the public retirement pensions, which in most European countries form the only support in the old age, convinced people to take additional private precautions. Due to the high returns and the relatively low risks of stocks in the long run, they seem to be the optimal mean to invest the funds allocated for this reason.

Many people have not the knowledge to invest directly into the stock market and therefore use mutual funds for their investments. This is why a large number of new mutual funds have emerged in recent years and raised huge amounts of money. Increased competition between mutual funds and other institutional investors required them to show a high performance in every year. They tried to achieve this goal through means of active management, i.e. making profits on short term movements in the stock market by frequent trading. Consequently, trading volume increased significantly. On the NYSE trading volume has grown 20 fold since the early 1980s and on the NASDAQ 60 fold.

Recently, advanced telecommunications technologies, like the internet, enabled also private investors to exploit these short term movements of the stock market, which is called *day trading*, increasing trading volume further.

Nowadays stock markets directly compete with each other for trading volume and established exchanges are challenged by newly created electronic trading platforms. This more competitive environment has made it crucial for stock markets to offer a trading environment that is attractive for investors, private as well as institutional.

1.1 The Christie-Schultz debate

On May 26, 1994 a study from *William G. Christie* and *Paul H. Schultz* became public, which showed that on the NASDAQ market makers quoted prices such that odd eighths were rarely found for about 70% of the most active traded stocks.[1] On

[1] This study was published in CHRISTIE AND SCHULTZ (1994).

the NYSE odd eighths quotes are also less frequent found than even eighths quotes, but to a much less extend than on the NASDAQ. As they also found the spread to be significantly higher on the NASDAQ than on the NYSE, they attributed this finding to market makers implicitly colluding to quote wider spreads.[2]

These findings gave way to investigations by the SEC as well as the Department of Justice into the causes of this behavior. Only in February 1999 the investigation has been terminated with an settlement out of court. The NASDAQ reacted promptly to the findings by changing their order handling rules in late 1994 and again in 1997.

In a companion article, CHRISTIE ET AL. (1994), it is further reported that only a few days after these results have become public, not only the spread of the affected stocks reduced significantly, but also that odd eighths quotes have been used much more frequently. As no evidence of a change in the costs could be found, this result further supported the suspect that market makers had been colluding.

In the aftermath of the publication of these two articles, a broad discussion whether NASDAQ market makers had been colluding, emerged. This discussion is also called the *Christie-Schultz debate*. A large number of empirical investigations seem to support the collusion hypothesis while others offer different explanations of these findings.[3] Theoretical contributions addressing implicit collusion have been rare thus far. The only theoretical paper addressing the problem directly is DUTTA AND

[2] The bid-ask spread is the difference between the highest price at which an investor can sell the asset, the bid price, and the lowest price at which he can buy it, the ask price. The bid-ask spread is a source of revenue for market makers. By selling the asset at a higher and buying at a lower price, they can make profits. On the other hand, the spread imposes a cost on traders as they frequently buy and sell the stock within a very short period of time. Large differences between the price to buy and the price to sell the stock, i.e. large spreads, make it very difficult to make profits from small price fluctuations.

[3] For example, FURBUSH AND SMITH (1996) pointed out that those assets not using odd eighths experienced larger absolute spreads, whereas the relative spreads to their price were found not to be higher. Their results only give evidence of clustering,clustering but not of collusion. KLEIDON AND WILLIG (1995) suspect that two competitive equilibria exist. One equilibrium uses a wide spread and offers a large depth, while the other equilibrium applies a narrow spread with a small depth. In light of the above described findings and the public pressure on market makers, they suggest that market makers changed their behavior such that instead of the first the second equilibrium had been applied. The observed odd eighths avoidance in their view is only the result of quote clustering not affecting trading costs.

MADHAVAN (1997). Their model shows under very restrictive assumptions that market makers can collude to quote a larger spread and make extraordinary profits.

As we will see below, the spread directly affects the competitiveness of an exchange. Therefore it is useful to investigate the possibility for market makers to collude implicitly on wide spreads, how to find direct evidence of implicit collusion and which measures are appropriate to reduce the effects arising from implicit collusion. We will address these issues in this work.

1.2 The importance of transaction costs

The attractiveness of an exchange or any other market place is called *market performance*. As the market performance cannot be determined directly, it is necessary to develop indicators. MARQUARDT (1998) identifies several measures to determine the performance of financial markets. According to his analysis the most important measures are the informational efficiency of prices, fast execution of orders, and transaction costs. BROWN (2000) additionally mentions market transparency, i.e. the disclosure of trade information.

An empirical investigation by ECONOMIDES AND SCHWARTZ (1995) into the behavior of professional traders shows that transaction costs are a major concern in trading. They found that lower transaction costs are one of the most important motives for traders to use electronic trading platforms instead of an established exchange. Most traders are willing to accept a certain delay in the execution of their orders if they can reduce their transaction costs. Informational efficiency of prices is also not very important for professional traders. Their short time horizon forces them to focus on the price changes within the next few hours rather than the long-term fundamental value of an asset. On the other hand, market transparency is vital to them as they more easily can identify short-term trends from recent trades and unexecuted orders.

According to BROWN (2000), transaction costs can be divided into explicit and implicit transaction costs. *Explicit transaction costs*, i.e. exchange fees, broker commissions and taxes, have decreased substantially in the last years due to liberalizations, increased competition and more advanced computer technologies. These costs are only of concern for a minor part of the traders. *Implicit transaction costs*, the bid-ask spread and the market impact of orders[4], however, are a major concern for most traders.

In the light of these results, markets should focus on the reduction of implicit transaction costs. The spread is the most important determinant of transaction costs at all. LESMOND ET AL. (1999) report a correlation of 88% between the spread and transaction costs, such that we can reasonably focus on the spread as the major source of transaction costs and hence as a measure of market performance.[5]

The spread and market impact, especially depth, are closely related, as we will see in chapters 4.3 and 6.7, therefore it is not possible to address them independently. Despite the importance of transaction costs, on which we concentrate in this work, other measures for market performance, informational efficiency, fast execution of trades and transparency should not entirely be neglected as they indirectly also affect transaction costs.

In this work we want to figure out whether implicit collusion between market makers is sustainable and how the spread is affected under much more general conditions than in the model of DUTTA AND MADHAVAN (1997). Based on these results and empirical investigations we will finally present some measures to reduce the effect arising from implicit collusion.

[4] The market impact describes the influence a certain order has on prices in the market. A part of the market impact is the depth of the market, i.e. the maximum order size that can be submitted such that it is executed entirely at the currently displayed price.

[5] MARQUARDT (1998) also identifies the spread as the most important measure of market performance following more theoretical considerations.

1.3 Framework of the analysis

Neoclassical economic theory assumes that all trades between individuals are costless and take place in an instant. The market coordinates individual decisions through the aggregation of demand and supply and the equilibrium price is determined such that they are exactly balanced. In the market buyers and sellers are assumed to interact directly without any intermediary.

In reality, however, we see a completely different picture. Most individuals do not directly buy their goods from the company that produces them. For example, wholesale and retail companies buy the goods from producers and sell them to consumers; loans are in most cases not directly granted by individuals investing money, instead banks and other financial institutions receive their investments in form of bank deposits and use these proceedings to grant loans. Obviously, in most cases intermediaires can be found that link buyers and sellers. SPULBER (1999, p. 22) reports that about 25% of the GDP in the United States is generated by these intermediaries.

The creation of intermediaries lowers transaction costs that arise from the interaction of consumers and producers. In recent years the impact these intermediaries have on the economy, has more and more been taken into account and given rise to a new field of academic research, called *market microstructure*. SPULBER (1999) gives an overview of this entire field.

Intermediaries can also be found in financial markets. They are specialized market participants like brokers[6], market makers[7], and match makers[8]. These market participants are subject to certain rules that each exchange and national legislation imposes on them, called the *market structure*. The market structure affects the behavior of all market participants and hence also transaction costs for investors.

[6] Brokers collect orders to buy or sell an asset from their customers (investors) and route them to the exchange.

[7] Market makers have the obligation to buy and sell an asset upon demand from brokers at prices they have to publish prior.

[8] Match makers collect the orders arriving from brokers and execute them with each other.

In this work we will restrict our attention to the analysis of the behavior of market makers and put a special emphasis on their price setting behavior, i.e. how they determine the bid and ask prices they quote and which spreads emerge.

In market microstructure theory models have been developed to determine the costs of market makers and the associated price setting behavior. Thus far, most models assumed the costs of market making to form the basis of their behavior. We will extend this analysis by taking into account that market makers act on self-interest, i.e. maximize their expected utility. As they are operating in the same market over a long period of time under similar conditions, it seems to be natural assuming them to behave strategically in order to maximize the present value of expected future utility. This will be taken into account by applying models from game theory to determine their optimal behavior and the implications for the bid and ask prices.

Throughout this work we will make the standard assumptions of most economic models: all market participants behave rationally.[9] The models employed are static throughout. We are primarily interested in the determination of equilibria and its properties rather than the dynamics how these equilibria are achieved. As only the behavior of market makers is considered, we focus on partial equilibria. How the behavior of market makers affects decisions of investors, brokers, consumers, producers and so forth is not explicitly taken into account.

1.4 Organization of the work

In *chapter 2* we will give a short overview of the trading process in stock markets. The observed market forms and their advantages and disadvantages will be introduced and the role of different market participants is pointed out. Then various forms of orders and rules for their execution are described. Finally, the impact of electronic trading and some more recent developments in stock markets will be considered.

[9] In Appendix C we will point out some objectives brought forward by critics of this approach.

The motivation for this work were the findings of implicit collusion between market makers on the NASDAQ. As also most empirical work focuses on the NASDAQ, we give a short introduction of the rules applied by this market in *chapter 3*. Besides the trading rules we also consider its history, organizational structure, conditions for companies to list their shares and prerequisites to have direct access to the market as broker or market maker.

Chapter 4 will give an overview of the most important contributions in market microstructure theory. At first we will present models showing that intermediaries facilitate trading in financial markets and which form of intermediation, market or match making, is preferred. After this in *chapter 4.2* auction markets will be considered. It will be shown that information is gradually incorporated into prices through trading. Thereafter we concentrate on dealer markets. In *chapter 4.3* we assume risk averse market makers trading with equally informed investors. We will show that in this environment market makers face inventory costs, which forces them to quote a bid-ask spread. It turns out that the spread and market depth are positively related. Assuming risk-neutral market makers trading with investors having superior information, causes market makers to bear adverse selection costs, which also causes them to quote a bid-ask spread, as will be shown in *chapter 4.4*. However, they face no inventory costs in this case. We will see that the spread and trading volume are positively correlated. The implications of including limit orders are to be considered in *chapter 4.5*. As thus far only markets with a single asset have been given attention, *chapter 4.6* investigates any implications that arise when introducing markets with multiple assets. Finally, *chapter 4.7* reviews some empirical investigations into the importance of the found costs, order processing costs, inventory costs and adverse selection costs for the bid-ask spread.

As the current literature on market microstructure in most cases focuses on the costs of market making, we will investigate the price setting behavior in more detail in *chapter 5* using a game theoretic framework. For this reason *chapter 5.1* reviews the most important elements from game theory used in this work. *Chapter 5.2* intro-

duces the basic model used in the further analysis and derives the competitive and cooperative equilibria as benchmarks. In *Chapter 5.3* we then investigate strategic behavior of market makers in the price setting behavior. Starting with the Dutta-Madhavan model we show that strategic behavior can give rise to market makers quoting noncompetitive bid-ask spreads generating excess profits. A prerequisite is that they discount future profits not too much, i.e. are sufficiently patient. We then will generalize this model by introducing different costs of market making, imperfect knowledge of relevant parameters and imperfect observation of other market makers' behavior. In all these cases we find that implicit collusion is possible for sufficiently patient market makers. However, we will also see that with these generalizations empirical evidence will be more difficult to find.

After showing that order preferencing facilitates implicit collusion in *chapter 5.4*, we derive in *chapter 5.5* that implicit collusion is further facilitated for market makers being competitors in several assets. Thereafter we demonstrate that market makers can also implicitly collude in determining the assets they act as market makers for, such that indirect competition between assets is reduced. As the models presented in this chapter are not only difficult to implement for market makers, but empirical evidence will also be difficult to find, we finally in *chapter 5.6* present a method how market makers can better coordinate their decisions and sustain implicit collusion more easily by using focal points. Empirical evidence from the literature as well as an own investigation will be presented to support the theories developed in this chapter.

In *chapter 6* we will consider several measures to limit the extent of implicit collusion and hence reduce the bid-ask spread for investors. We find that the most appropriate measures are to enhance outside competition from limit orders and time priority of the quotes set by market makers. To a less extent also anonymity of the market makers posting quotes and increased market transparency may reduce the bid-ask spread. Arguments from theoretical models as well as empirical investigations described in the literature are used to give evidence on the effectiveness of the proposed

measures.

In the *concluding remarks* we summarize our findings and point out several other aspects that are worth to be considered in future research to reduce the trading costs of investors.

The *appendices* provide additional information that could not be included in the main body. In *appendix A* we give an introduction into the utility theory most common to apply in economics, expected utility. *Appendix B* presents the portfolio selection theory, which is used in deriving inventory costs in chapter 4.3. An introduction to rational expectations and Bayesian learning in *appendix C* is used in chapter 4.4 to determine the adverse selection costs. Some mathematical terms not common to use in economics, but needed in chapter 5, are defined in *appendix D*. To limit the space in the main text, we present longer proofs of results from chapter 5 in *appendix E*.

Chapter 2

The organization of trading in stock markets

This chapter will give a brief overview of the organization of stock markets. The aim is not to give a survey of all possible regulations and special forms found at different stock exchanges, but to explain the basic principles.[1] A special focus is laid on those aspects that have an impact on the price setting in these markets.[2]

After some preliminary definitions in chapter 2.1, the importance of stock markets for investors, companies and the society as a whole will be pointed out in chapter 2.2; chapter 2.3 describes the different forms of trading on an exchange and chapter 2.4 shows how orders are submitted to an exchange, who participates in trading and what forms of orders exist; according to which rules orders are executed is explained in chapter 2.5; chapter 2.6 deals with the influence of computerization on the organization of stock markets and trading; finally, chapter 2.7 points out some recent developments in financial markets.

[1] LEE (1998) gives a very detailed and theoretically based overview of the current regulation of stock markets in the United States, including recent developments in electronic trading.

[2] An overview of the regulation of stock markets is given in chapter 3 for the NASDAQ and in SCHWARTZ (1993, ch. 2 and 4) for the NYSE and the London International Stock Exchange. A detailed description of the regulations in the United Kingdom, United States, Switzerland, France, Netherlands, Austria, and Japan can be found in HOPT ET AL. (1997, Part III) and in HOPT AND BAUM (1997) for Germany. A history of stock exchange regulations in Europe is given in MERKT (1997).

2.1 Definitions

A *market* is a place where supply and demand for a good meet,[3] i.e. potential sellers and buyers of a good submit their supply and demand schedules. This place has not to be a certain location,

> "... but the entire territory of which the parts are united by the relations of unrestricted commerce that prices there take the same level throughout with ease and rapidity."[4]

To achieve this tendency towards an equal price, JEVONS (1911) emphasizes the importance of communication between market participants, i.e. the transmission of information, especially prices:

> "The traders may be spread over a whole town, or region of country, and yet make a market, if they are (...) in close communication with each other."[5]

MARSHALL states the characteristics a good must have to be traded in a market:[6]

- The good has to be *widely demanded*. A good that is not widely demanded cannot be traded frequently, hence there is only infrequent intercourse and too few prices that are needed to form a market as stated in the above definition of Cournot.
- The good has to be *homogeneous*.[7] If a good is not homogeneous, e.g. pieces of art or special machines, the pieces are not interchangeable with each other,

[3] See SCHUMANN (1992, p. 23).
[4] COURNOT (1838, p. 42).
[5] JEVONS (1911, p. 81).
[6] See MARSHALL (1920, p. 141f.).
[7] Goods are homogeneous if they are interchangeable, i.e. one piece of a good has exactly the same characteristics as another.

hence they will fail to have a tendency towards equal prices as needed by the definition of a market.

- The good has to be *transferable* and *storable* at low costs, compared to its value. High costs of transferring the rights or storing the good would hinder the free intercourse wanted by Cournot to form a market. The gains that have to be made from trading the good to offset these costs would be too high.

Technological innovations reduced the costs of transfer and storage significantly in recent years, thereby increasing the number of goods that can be traded in markets. New information technologies enable traders to communicate at low costs over long distances, enlarging the region that can be viewed as a market. Further improvements in standardization made more goods homogeneous, so that they can be traded in markets nowadays. The reduction in costs and standardization, besides further improvements, also increased the demand for many goods. Hence today more and more goods can be and are traded in markets.

Securities are rights that are chartered in a document. To execute or transfer this right the document has to be presented.[8] *Financial securities* are securities, where the rights are a sequence of future cash flows.[9] The future cash flow can consist of money (interest, dividends) or other financial securities (e.g. stocks or bonds). Most securities fit into one of the following categories: stocks, bonds, derivatives[10] or a combination of these. When addressing securities in economics, financial securities are principally being referred to. We will stick to this tradition in the remaining parts of this chapter.[11]

Securities are typically divided into smaller parts, each part representing a fraction of the whole. Therewith securities are homogeneous. They are transferable at very low

[8] See Büschgen (1991, p. 782).

[9] See Dumas and Allaz (1996, p. 1).

[10] A derivative is a security whose future cash flow depends on the value of another security.

[11] Another frequently used expression for financial securities is *asset*. In relation with prices and valuation this term is the most frequently used, while in relation with market regulations the term securities is more common. Therefore, starting in chapter 4, we will use the expression *asset* for securities.

costs and storage costs can be neglected.[12] As securities are also widely demanded as a mean for investments, they fulfill all characteristics to be traded in a market.[13] These markets are called *security markets*.[14]

We have to distinguish two forms of security markets: primary and secondary markets. In *primary markets* new shares of a security are issued, i.e. for the first time sold to an investor,[15] while in *secondary markets* already issued shares of securities are traded among investors.[16]

If the amount of an existing security is increased, the new shares can be issued either by organizing a separate auction or by selling them in the open market, i.e. they are issued by placing a sell order from the issuer in the secondary market. In this case the issuer behaves like an investor. Although this form of issuing formally belongs to the primary market, it is referred to as an operation in the secondary market. Sometimes this form of increasing the amount of an existing security is also applied for issuing a new security, especially in OTC markets for derivatives.

In futures and options exchanges there does not exist a secondary market. All transactions take place in the primary market according to the definition above. If an investor buys a derivative, another investor has to issue a new share of this security, an existing derivative cannot be bought. The outstanding amount of these derivatives is not fixed as in the case of stocks or bonds. If an investor wants to sell a derivative he has bought, he issues a derivative which exactly offsets the derivative he wants to sell.[17] Formally, all these operations must be placed into the category of

[12] Many securities do not exist physically nowadays and if they do, they are stored at a clearing house, imposing very low storage costs. Formerly securities had to be handed over, but today the transfer usually is only booked into accounts kept at the clearing house.

[13] There are sometimes securities issued that are not widely demanded. They typically vanish after a short period of time.

[14] Often the term *financial markets* also refers only to security markets, although they in general encompass also rights that are no securities, such as bank loans or deposits. Often foreign exchange markets and even the commodity markets are included into the term financial markets, as the characteristics of these markets are very similar.

[15] We call investors all those, who are invested in securities or are interested in being so (potential investors).

[16] See SCHWARTZ (1988, p. 3).

[17] For every derivative such an offsetting derivative exists because it can be found in two forms: as a long and as a short position. An investor is long if he has bought a security, he is short if

primary markets. But since they have all characteristics of a typical trading activity, they are assigned to secondary markets.

In most cases, security markets are being referred to the category of secondary markets. This convention will also be applied in this text. The remaining analysis will be concentrated on stock markets, but most concepts and findings can easily be adapted to other security markets.

2.2 The importance of organized stock markets

In most cases the time horizons of investors and the issuer of a security do not coincide. While companies have a need for capital of very long, even infinite disposability, investors on the other hand may want to change their investments to adjust for new information, changed tastes, or liquidity needs. As the amount of outstanding shares is fixed (disregarding periodic increases and repurchases of capital) all shares have to be held by investors. To adjust their investments, investors have to trade them with each other. KEYNES (1936, p. 151) pointed this aspect out as follows:

> "In the absence of security markets, there is no object in frequently attempting to revalue an investment to which we are committed. But the Stock Exchange revalues many investments every day and the revaluations give a frequent opportunity to the individual (though not the community as a whole) to revise his commitments."

If no securities markets exist, it is very difficult to find another investor who wants to take the counterpart in a trade. The lack of communication between investors makes

has issued the security. In stock and bond markets investors usually are long and the company who has issued the security is short. However, there is the possibility of short sales by investors in many markets, so that only on average investors have to hold a long position. Adding a short and a long position of the same security exactly offsets the investor.

the search for a counterpart very costly.[18] When a counterpart is finally found (e.g. by advertising) the next problem is to find a price at which both are willing to trade. The valuation of a stock depends on the information an investor has, inconsistent information will make it difficult to find a suitable price.[19] Even if both can agree on a price, the trade may occur at a "wrong", i.e. informationally inefficient, price[20] when neither participant has precise information. This gives wrong incentives to investors, resulting in an inefficient allocation of resources.

To overcome these difficulties markets provide investors with two services: liquidity and the aggregation and revelation of information.

By providing *liquidity*, markets facilitate the exchange of assets between investors. As investors meet on markets, the costs of searching a counterpart for a trade is reduced significantly, a counterpart can easily be found by addressing the market. Competition between investors for a trade will further ensure that a better, informationally more efficient price will be charged. These reduced costs of trading will allow investors to adjust their investment decisions more frequently to their information and tastes, resulting in a more efficient allocation of resources.[21]

As markets generate prices, they can be used to reveal and *aggregate information* on a security.[22] Compared to other sources of information, prices can be observed at nearly no costs. Without having much additional costs, an investor can increase

[18] Chapter 4.1.2 presents a formal model showing how these costs can give rise to the formation of organized markets.

[19] An overview of asset valuation can be found in nearly all textbooks on finance. Good synopsis are INGERSOLL (1987), DUFFIE (1996) and COCHRANE (2000), besides others.

[20] Prices are called *informationally efficient* if they fully reflect all available information. FAMA (1970) distinguishes three forms of efficiency, depending on the information available: weak, semi-strong and strong efficiency. Prices are *weakly efficient* if they reflect only information derived from previous prices or returns, if all publicly available information is reflected in the price, it is called *semi-strong efficient*. In cases where also private information available to any market participant is reflected, prices exhibit *strong efficiency*.

[21] ARROW (1964) stresses the importance of asset markets for an efficient allocation of risks between individuals.

[22] It was HAYEK (1945) to point out the importance of prices as a source of information. How prices can aggregate and reveal information to investors is discussed in more detail in chapters 4.2 and 4.4.1.

his information and in this way reduce the risk to trade at a disadvantageous price resulting from a lack of information. This further reduces his costs of trading.

By holding a portfolio that fits better his tastes and information, an investor reaches a higher level of utility. The reduced costs of adjusting his investment decisions increase his returns and hence the price he is willing to pay for an asset. This benefits also the issuers of assets as they can issue their assets at higher prices, reducing their costs of capital and increasing profits.[23] Increased profits give incentives for further investments and hence promote economic growth.

From the existence of security markets investors profit from higher returns and a better allocated portfolio, issuers of securities from lower costs of capital and higher profits and the society as a whole from a more efficient allocation of resources, higher investments and growth. Therefore everyone benefits from the existence of securities markets.[24]

A further reduction in the costs of trading can be achieved by applying uniform rules to the trading process. Security markets that apply a fixed set of rules governing trading are called *organized* security markets or an *exchange*.[25] The rules of an exchange should regulate

- The *admission* of stocks. The stocks to be traded on an exchange have to meet certain standards in order to guarantee a minimum of investor protection and ensure regular trading. These standards differ very much between exchanges.
- The *access* to the market. It has to be determined who is allowed to trade directly in the market and what conditions have to be met for this access. By applying certain standards, the risk of a counterpart failing to fulfill his duties after a trade has been negotiated, can be reduced.

[23] See KEYNES (1930, Vol. II, p. 195).
[24] See SCHWARTZ (1988, pp. 4 ff.).
[25] A more formal definition of an exchange, based on the current legislation can be found in LEE (1998). RUDOLPH AND RÖHRL (1997, p. 168) also apply a similar definition of an exchange and point out that recent innovations, especially the computerization of exchanges (see also chapter 2.6), made it necessary to adapt the traditional legal definition of an exchange.

- The *types of orders* that can be submitted. A standardization of the order types simplifies the trading process and the set of rules can be held small.

- The *execution* of orders. Rules on the execution of an order include the rules to determine the prices at which a trade occurs and when trades are executed. These rules can avoid the problem of some investors gaining the advantage from the cost of others, e.g. as a consequence of personal links to other market participants.

- The *clearing* of trades. A standardization of the clearing process, i.e. the way and time trades are settled, avoids a separate negotiation on this point. Most exchanges do not only set rules for the clearing process, but organize it by themselves through special clearing houses.

Besides the reduction of trading costs, the standardization of trading also makes prices more comparable to each other, both over time and between assets. The prices therewith can better aggregate and reveal information.

The rules imposed on the trading process are called the *market structure*.[26] They can be imposed by public regulation or self-regulation of the market. In most cases a combination of public regulation and self-regulation can be found.

The structure of an exchange will influence the costs of trading, which in turn will affect the price formation. We will therefore investigate the process of price formation in more detail in this chapter by giving a short overview of the basic market forms.

2.3 The different market forms

The *market form* describes the way trades occur in a market. In standard neoclassical theory the market form implicitly assumed is that of an *Walrasian auctioneer*.

[26] See O'HARA (1995, p. 1).

The auctioneer suggests a price to the investors. The investors[27] determine which amount they are willing to buy or to sell at the stated price and submit their decisions, called *orders*, to the auctioneer. If demand and supply do not exactly equal, no trade occurs. The auctioneer suggests a new price and the investors revise their decisions. This process continues until aggregate demand exactly equals aggregate supply. The price at which aggregate demand and supply are balanced, is called the *equilibrium price* or *market clearing price*. If the equilibrium price is found, all orders are executed in a single multilateral trade.[28]

This process mostly is assumed to be finished in an instant of time, imposing no costs on investors. In reality, however, the revision of orders and the announcement of a new price would take considerable time and impose high costs, especially if many investors are trading on the market. Furthermore, it would be pure incident if a price could be found that clears the market exactly. Both, prices and quantities are discrete, either by definition as no fractions of a stock can be traded or by convention, e.g. discrete prices. Additional rules have to be applied to determine the price with these small imbalances.

For these reasons a Walrasian auctioneer seldom exists in real markets, other market forms have been established over time.[29] There exists a wide variety of market forms around the globe, every exchange has its unique way of executing a trade. Also within an exchange there often exists different market segments, each having its own rules.[30] Despite these differences in detail, the market forms can be organized in one of six main market forms. Figure 2.1 shows the classification of market forms

[27] In this section use the term *investor* for all market participants and assume that they directly interact at the exchange. However, in reality most investors have to use an agent for trading. This part of the structure of an exchange will further be discussed in the following sections.

[28] See O'HARA (1995, p. 4).

[29] The only examples of markets that use the concept of a Walrasian auctioneer are the London Gold Fixings at 10.30 am and 3 pm and the Frankfurt Foreign Exchange Fixing for selected currencies at 12 am that has been ceased to exist after December 30, 1998 with the introduction of the EURO. See O'HARA (1995, p. 7).

[30] Different rules can often be found for frequently and infrequently traded stocks. The Frankfurt Stock Exchange with its official quotation (Amtliche Notierung), regulated unofficial market (Freiverkehr) and Neuer Markt is a good example for an exchange with different market segments applying different sets of rules.

as will be used below.

In all market forms orders[31] can be submitted to the market at any time, what differs is the time and way these orders are executed. A market form which is very close to the concept of an Walrasian auctioneer is the *batch system* or (*periodic*) *call market*. In batch systems incoming orders are not executed immediately, but stored and executed in a multilateral trade at a predetermined point of time. The price that will be applied for this trade, is the price at which most orders can be executed, i.e. the price with the highest trading volume. We find batch systems in two forms: à la criée and par cassier.

Trading *à la criée* allows investors to revise their orders until the time of execution. To give additional information for the revision of orders, the price that would be applied if all orders were to be executed immediately, is continuously published.[32]

Trading *par cassier* allows investors not to revise their orders and typically the price that would be applied in case of immediate execution of all orders ,is not published.[33]

Batch trading can rarely be found in regular markets. It is only frequently used to determine the opening and sometimes the closing price of a trading day and to determine the price of some infrequently traded stocks.[34]

In *continuous markets* trades can occur not only at predetermined points of time, as in batch trading, but at any time two orders can be executed. For every submitted order it is immediately checked whether there exists another order on the market,

[31] We find two forms of orders: orders that specify the worst price (highest price for a buy order, lowest price for a sell order) at which they can be executed (limit orders) and orders to be executed at any price (market orders). These order forms are presented in more detail in section 2.3.

[32] See COHEN ET AL. (1986, p. 16). The term *à la criée* refers to the verbal order submission to the auctioneer that has been used at the Paris Stock Exchange prior to the introduction of electronic trading in 1986. However, it is not relevant whether the order is submitted verbal or written, the important feature of this market form is the publication of the price and the possibility to revise orders. The name is only kept for historical reasons.

[33] See COHEN ET AL. (1986, p. 17). Like à la criée the term *par cassier* is kept for historical reasons, as in this market form orders were submitted written. Important is only the impossibility to revise orders.

[34] See also table 2.1.

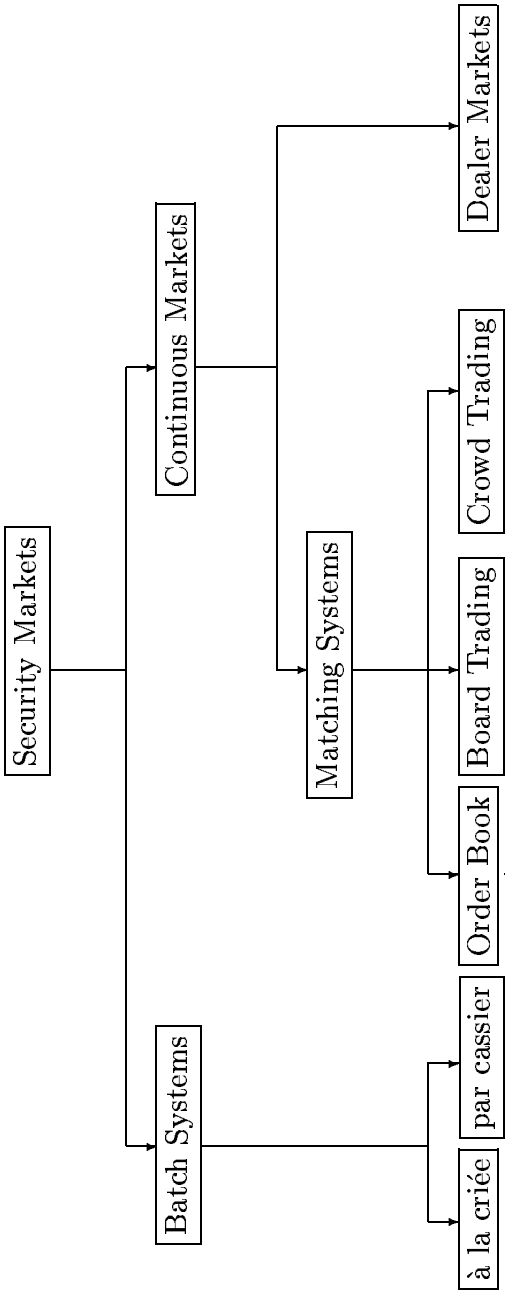


Figure 2.1: Classification of market forms

such that these orders can be executed in a bilateral trade. If no such order exists, the order is stored and executed with the next matching order arriving on the market. We find two forms of continuous markets: dealer markets and matching systems.

In *dealer markets* special market participants, called *dealers*, have the obligation to "make the market", hence they are also called *market makers*. Every market maker has publicly to set prices at which he is willing to sell (*ask price*) and to buy (*bid price*) the security.[35,36] The bid and the ask prices have not to be, and will not be, equal, as we will see in chapter 4. At the stated price the market maker has to sell (buy) the security immediately from (to) any investor demanding this. The market maker trades on his own account, i.e. he forms the counterpart of the investor.

In *matching systems* no special market participants exist to form the counterpart by trading on their own account. The trades are only bilaterally executed between two investors. Three different forms of matching systems are known: order book systems, board trading and crowd trading.

In *order book systems* all submitted orders are stored in an order book. If two orders cross, they are immediately executed and the price and volume of the trade are published. In some cases the order book keeper also acts as a market maker, like on the NYSE. A further distinction can be made whether the order book is *open* or *closed* to the public, i.e. if the investors can look into the order book or not. Mixed forms can also be found where the order book keeper can give some information, as on the Frankfurt Stock Exchange.

In *board trading* the prices at which investors are willing to trade are also entered into an order book, but only the best (highest) bid and the best (lowest) ask prices are published on the board, e.g. the two best prices on the Hong Kong Stock Exchange. If an order arrives on the market and accepts the best price stated on the board, they are immediately executed. The price at which the trade takes place is published

[35] When setting these prices, the market maker does not know whether the next order arriving at the market is a buy or a sell order, what its size is and when an order will arrive.

[36] The prices a market maker sets are also called *quotes* or *quoted prices*.

on a separate board and the executed orders are cancelled from the order book. In this system the trade size is fixed to what is called a *lot*. This facilitates trading because the size of an order has not to be considered in matching them.

The third form of matching systems is *crowd trading*. Investors meet on the trading floor of the exchange and discuss the prices at which they are willing to conduct trades. If two investors agree upon a trade, their orders are executed and the price is published.[37]

In another classification matching and batch systems are also called *auction markets*, to distinguish in this classification between matching and batch systems, matching systems are called *continuous auctions*.

As can be seen from table 2.1 every market form can be found on least at one of the leading stock exchanges.[38] There is no dominating market form to be found, what suggests that every market form has its advantages and disadvantages, although currently we observe a tendency towards order book systems.

Batch systems have the advantage of collecting orders over a longer period of time. Large order imbalances that may occur over time, e.g. a large order arriving on the market, will have a smaller effect on prices than with an immediate execution of the order. Often an imbalance is reduced over time and the volatility of prices diminishes. If the trading is à la criée, investors can react to this imbalance as they can observe that the price would change significantly with execution. On the other hand, to revise an order imposes not only costs on investors, but also on the exchange. The order flow becomes difficult to handle and errors are more likely to occur than in trading par cassier. On many stock exchanges with matching systems all orders accumulated over night are cleared immediately at the beginning of the trading session in one multilateral trade at a single price. This is advantageous

[37] See COHEN ET AL. (1986, pp. 19 ff.) and SCHWARTZ (1988, pp. 24 ff.).

[38] Only crowd trading cannot be found since the Swiss Exchange Zürich changed to an electronic trading platform in 1996. Of the more important exchanges nowadays only the London Metal Exchange applies crowd trading, but it is also planned to introduce electronic trading in the near future and hence change the market structure.

	Batch Systems		Continuous Markets			
			Matching Systems			
	à la criée	par cassier	Order Book	Board Trading	Crowd Trading	Dealer Markets
Frankfurt Stock Exchange						
Amtliche Notierung		opening	closed			
Freiverkehr		X				
Neuer Markt			closed			(X)
Hong Kong Stock Exchange				X		
London Stock Exchange						X
NASDAQ						X
NYSE		opening	closed			X
Tokyo Stock Exchange	opening		open			
Swiss Exchange Zürich			X			
Paris Stock Exchange			X			
Vienna Stock Exchange						X

Table 2.1: Market forms of selected stock exchanges

for the determination of opening prices. If they had to be executed in subsequent bilateral trades, these orders would hinder the handling of orders submitted at the beginning of the trading session.[39]

On the other hand, batch systems have the disadvantage that trades occur only a few times per day (once or twice). Investors have to wait a considerable time until they are able to trade the next time. The reaction to new information is not possible immediately, imposing waiting costs on investors.[40] As less prices are available to investors, the aggregation and revelation of information through the price system cannot be assured as good as in continuous markets.[41]

The advantage of faster execution of an order and therewith reduced waiting costs is one of the main arguments for continuous markets. In dealer markets the market maker guarantees immediate execution of an order, but, as we will see in chapter 4, he will not provide this service for free. The fees charged by the market maker may counteract the advantages of this market form.

Matching systems impose no additional fees on investors[42], but therefore immediate execution is not guaranteed as a matching order has to be found before an order can be executed, hence investors may also face waiting costs. Especially for infrequently traded stocks, the time until an order can be executed may be very long. This makes the differences to batch systems in terms of waiting costs less important, whereas the volatility of prices will in general be higher, as an order may have a substantial effect on prices. This favors batch systems for infrequently traded stocks, whereas for frequently traded stocks the impact of not too large orders on the price can be neglected, favoring matching systems as the result of more frequent trading.

[39] See COHEN ET AL. (1986, pp. 23 ff.) and SCHWARTZ (1988, pp. 18 ff.). The NASDAQ has no call auction at the opening, hence trading volume is very high and the execution of orders submitted at the beginning of the trading hours takes a considerable time. Therefore they are currently considering to introduce a call auction at the opening.

[40] See GROSSMAN AND MILLER (1988).

[41] With trading à la criée prices are published continuously, but they will be biased up to short before the execution. In order to save costs, investors will not adjust their orders permanently, but only once just prior to the execution.

[42] We neglect here, as stated in the introduction, any direct fees levied by exchanges, brokers and any taxes to be paid.

Even for frequently traded stocks, where no considerable waiting costs exist, matching systems have the disadvantage that the price at which the trade will be executed, is not known in advance. It depends also on the order with which it is matched, whereas in dealer markets the investor knows the stated price of the market maker at which he will trade. This uncertainty on the price may favor dealer markets despite the fee market makers charge.[43]

Crowd systems allow only for small trading volumes. The time to negotiate the price does not allow for too many investors trading. But this system ensures the best price in continuous markets, as all orders compete directly for a trade. Board trading is capable of handling larger trading volumes and order book systems are more flexible to handle orders of different sizes. With the ability of handling larger volumes due to advanced computerization, the costs of conducting a trade decreases for the exchange and hence for investors.[44]

As can be seen from the above discussion, every market form has its advantages and disadvantages. The optimal market form depends on the nature of the stock, e.g. the frequency it is traded, and the personal tastes of the investors, e.g. their risk aversion or time preferences. A discussion of these aspects can be found in ANGEL (1997), where he suggests optimal trading rules for stocks of small companies.

In addition to the market forms presented above, many mixtures, like the NYSE, and variations of these pure forms exist. Special trading forms, e.g. off-exchange trades for large orders (block trading) or special trading facilities for small orders, complete the list of market forms.

2.4 Market participants and order submission

In the last sections it has been assumed for the sake of simplifying things that investors trade directly on the exchange. In reality, however, they have to use

[43] See GROSSMAN AND MILLER (1988).
[44] See COHEN ET AL. (1986, pp. 23 ff.) and SCHWARTZ (1988, pp. 18 ff.).

an agent, who trades for them on the exchange. This agent is called a *broker*.[45] A broker transmits the order he receives from an investor, his customer, to the exchange, where the order is treated according to the rules of the exchange. He does not trade on his own account.[46] The broker also informs the investor about the execution and the applied price of his order and settles the accounts. He also is liable for fulfilling the trade, hence the counterpart risk is reduced significantly.

In dealer markets the broker has to identify the market maker[47] offering the best price and transmit the order to this market maker for immediate execution. In matching and batch systems the broker only transmits the order to a *match maker*[48] and waits for its execution.[49] Figure 2.2 visualizes the way orders are submitted in different market forms.

Often the roles of market participants change. A market maker or broker may want to trade for his own account and hence by our definition become an investor, or on the NYSE the market maker also is keeper of the order book, i.e. match maker. Despite these mixtures of roles individual market participants have at a particular time, their activities will fit into one of the following categories: investor, broker, market maker or match maker.

The size of an order is not fixed, except in board trading. A market maker has to accept any order size at the stated price as long as the order is not too large. Large orders are normally divided into several smaller orders and executed over a longer time period, ranging between hours and months to avoid a significant influence on the price.

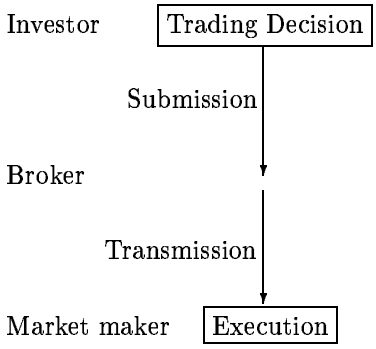
[45] There are special conditions that must be met to act as broker. In our context these conditions are of no interest and are therefore omitted here. Chapter 3.4 gives a detailed description of the conditions to be met for being granted access as broker to the NASDAQ.

[46] See O'HARA (1995, p. 8).

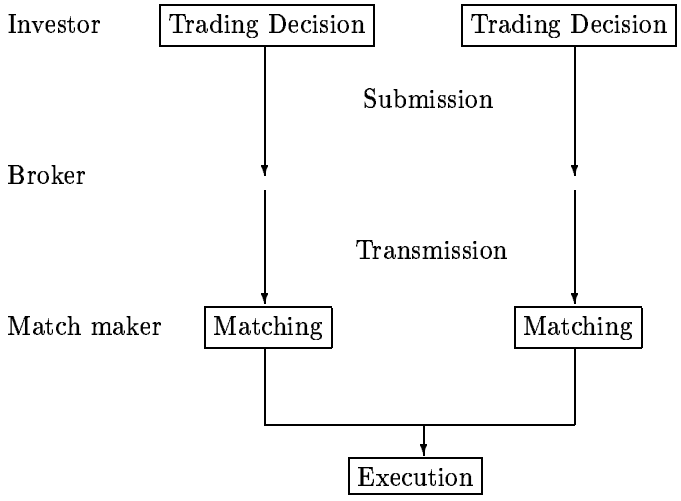
[47] To become a market maker very strict conditions have to be fulfilled. In some markets only one market maker per asset is allowed, e.g. the *specialists* at the NYSE. Chapter 3.4 gives more details on the admission as market maker to the NASDAQ.

[48] A match maker is a person that stores the order, i.e. keeps the order book, and initiates the matching of the orders.

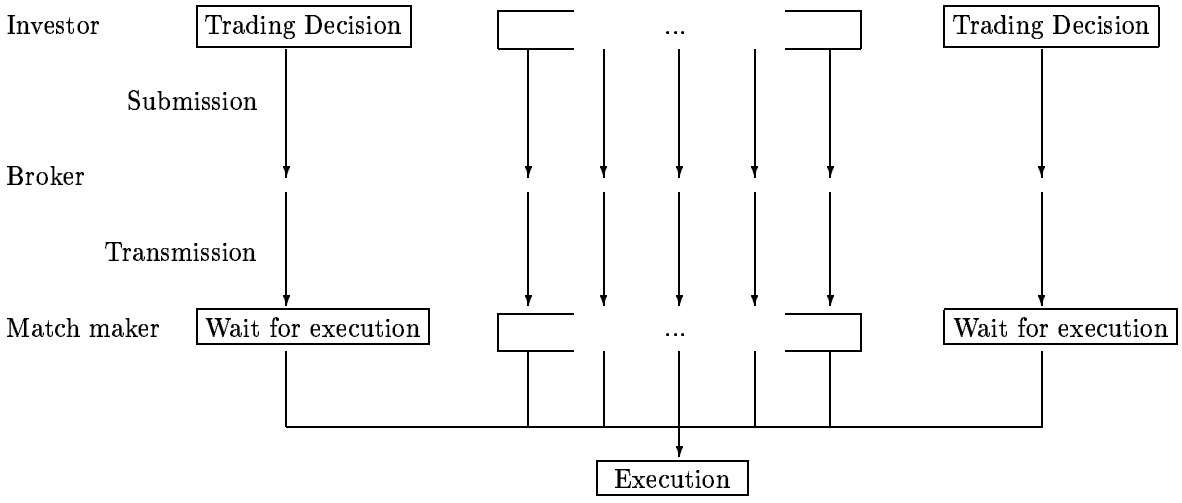
[49] See SCHWARTZ (1988, p. 18).



(a) Dealer markets



(b) Matching system



(c) Batch system

Figure 2.2: Order flow in different market forms

In most markets there exists a "normal" order size, as in board trading called a *lot*.[50] Orders with the size of a multiple of a lot can be divided into smaller, separate orders with the minimum size of one lot. These smaller orders will be executed separately with different matching orders or market makers at different points of time and prices. This splitting of large orders facilitates the execution of such an order and increases liquidity. If the orders were to be executed as a whole, it may be difficult to find a matching order with exact the same size.

Orders with a size smaller than one lot (so called *odd-lots*)[51] are sometimes traded by special market makers or match makers, respectively, or a batch system is introduced for these small orders. The Fixing on the Frankfurt Stock Exchange at 12.00 is an example for such a batch system.[52]

So far we only considered orders that were for execution at the best available price of the market. Such an order is called a *market order*. A market order will be executed at any price, in frequently traded stocks with continuous markets it normally will be executed within a short time after submission, as any offsetting order matches. In batch systems market orders are executed nearly with certainty.[53]

There exists another frequently used type of order, *limit order*. When submitting a limit order, the investor sets a maximum (minimum) price, the limit, at which he is willing to buy (sell) the security. Of course, he also will buy (sell) at lower (higher) prices. The execution of limit orders is not guaranteed, an offsetting order has to be found that fulfills these conditions. In most matching and batch systems the trade between two market orders is given priority over the trade of a limit and a market order. Therefore it can take a considerable time until a limit order can be

[50] The typical lot on the Hong Kong Stock Exchange is 5,000 shares. In 1998 the Frankfurt Stock Exchange changed its lot size from 100 shares to 1 share. On the NASDAQ the lot sizes vary between 100 and 1000 shares, depending on the stocks.

[51] Orders that are larger than one lot, but are not a multiple of a lot are split into a part consisting of a multiple of a lot and a part with the odd-lot. They are then treated as different orders. See KEENAN (1987, p. 23).

[52] See COHEN ET AL. (1986, pp. 27 ff.).

[53] There may be some special situations in which the execution is not guaranteed, but these situations have no practical relevance.

executed. Limit orders will be cancelled either at expiration, e.g. at the end of the day or upon withdrawal.[54]

There exist many other order forms, but they are rarely applied in security markets. One order form that has been important in the past is the *stop loss order*. Stop loss orders are executed as a market sell order if the market price falls below a certain limit, otherwise the order is not executed. This type of order has been used in the past, when an investor had not the possibility to observe the market continuously or contact his broker immediately. The aim was to prevent greater losses in these cases. Nowadays it is possible to receive recent news on security prices and communicate without problems nearly all over the world and stop loss orders can rarely be found.[55]

2.5 Trading priority rules

We can frequently run into the situation where there is more than one order unexecuted in the market which matches an incoming order. In this case it is important to establish rules deciding which of these orders will be executed, called *trading priority rules*. They not only have an impact on the time an investor has to wait until his order is executed and hence his waiting costs, but also on the price applied.[56]

The most important rule is *price priority*. With price priority market orders are executed first and only after all market orders have been executed, the limit order with the best price, i.e. lowest price for a buy order and highest price for a sell order, is executed, then the limit order with the second best price, and so forth.[57] This rule ensures that securities can be bought at the lowest and be sold at the highest

[54] See SCHWARTZ (1988, p. 17).

[55] See SCHWARTZ (1988, pp. 45 ff), where also an overview is given of other order forms that are possible to submit to US stock exchanges. However, these order forms are only rarely found and for this reason not further considered here.

[56] MOULIN (2000) provides an axiomatic treatment of priority rules. He shows that in each market there exists only a single optimal priority rule, however, his analysis does not allow him to determine this rule. DOMOWITZ (1993) gives an overview of the priority rules applied on several important stock markets.

[57] See COHEN ET AL. (1986, p. 156).

available price, reducing trading costs.[58] Price priority is found as the first priority rule at all stock exchanges.

The price priority rule will in many cases not be sufficient to distinguish between all unexecuted orders in the market. It will often be found that more than one order has been submitted at the same price or is a market order, so that we need additional rules for choosing the order that is executed with a matching order. These rules are called *secondary trading priority rules.*

The most common rule is *time priority.* An order that has been transmitted earlier by the broker is executed before an order transmitted later.[59] Another rule that frequently can be found, often in combination with time priority, is *size priority.* With size priority a larger order is executed before a smaller order. This rule in combination with time priority can be found at the NYSE[60] and since 1996 at the Toronto Stock Exchange, when the secondary priority rule was change from pure time priority.

In dealer markets the rule *public before dealer* is very important. If a limit order[61] submitted by an investor has the same limit as the price quoted by the market maker, the limit order is executed first.[62] Public before dealer is applied by the NASDAQ since 1997, while most other dealer market do not have this rule.

There are many other rules that have only minor importance in leading stock markets, such as *pro rata partial execution.* If at a certain price there is an imbalance in the orders, i.e. not all orders can be executed on one side, all orders on the larger side are only partially executed with the same fraction. This rule may be applied in

[58] See SCHWARTZ (1988, p. 18).

[59] See COHEN ET AL. (1986, p. 156). The time at which an order is transmitted, in most cases at which it is entered into the computer system of the exchange, is measured in hundredth of seconds to ensure a clear distinction between all orders, also in times of high volume. In dealer markets similar rules can be established to determine the market maker executing an incoming order.

[60] See SCHWARTZ (1988, p. 18) and COHEN ET AL. (1986, p. 156).

[61] Market orders have to be executed immediately either by the market maker or by a limit order. Consequently, there can be no unexecuted market orders.

[62] See COHEN ET AL. (1986, p. 157). If more than one limit order is unexecuted at this price they are distinguished by another rule, e.g. time priority.

batch systems. The least complicated rule is the *random selection* of the order that is executed.[63]

Not every rule can be applied in all market forms. There exists a wide variety of further rules, exceptions and modifications that are specific for a certain stock exchange.

The rules and market forms presented here form only a part of the market structure. Additional rules not yet mentioned encompass maximum price change limits, lower transaction costs for certain groups of market participants, besides others. An exhaustive description of all possibilities to form a market structure lies beyond the scope of this chapter.[64]

2.6 Electronic trading mechanisms

In recent years exchanges have computerized more and more functions. At the beginning of this process brokers only used computers to facilitate their order handling, e.g. the settlement of trades with their clients, the supervision of order execution and the clearing of trades. Exchanges used computers only for displaying and storing publicized data and for the clearing process. Orders had to be transmitted in conventional ways, i.e. verbal or written, from the broker to the exchange. Market makers and match makers as well as brokers had to be physically present on the trading floor of the exchange.[65]

Later the computer was used to assist market makers and match makers in handling the order flow by ordering the orders according to the priority trading rules and

[63] See COHEN ET AL. (1986, p. 157).

[64] Chapter 3 gives a more detailed overview of the market structure of the NASDAQ, including some, but not all, features omitted here. RUDOLPH AND RÖHRL (1997) give a more detailed overview of the economics and current state of stock exchange regulation.

[65] There existed also OTC markets that had no trading floor, but used a telephone for communication (telephone markets). On these markets in most cases only very infrequently traded stocks were listed, an exception has been the NASDAQ.

displaying the relevant information. The orders had to be entered into the computers by the market makers and match makers themselves.[66]

In 1973 the Frankfurt Stock Exchange had developed the first *automated order handling system*. It allowed brokers to transmit their orders electronically from their internal computer systems into the computer system of the exchange. The order was then automatically routed to the appropriate match maker. The execution of the orders still had to be done manually. Although the system has never been introduced in Frankfurt, other exchanges developed similar systems. These systems, e.g. the *DOT*-System of the NYSE implemented in 1976, were at the beginning only able to handle a small number of orders. For this reason the use was restricted to small orders. With time computer systems were able to handle more orders and the use was extended.[67]

The last step towards a *fully automated trading system* (or *electronic exchange*) was first taken in 1977 with the introduction of *CATS* on the Toronto Stock Exchange. It enabled not only the electronic transmission of orders, but without any interference was able to execute orders.[68] Like at the time automated order handling systems were introduced, these systems were at the beginning only able to handle small amounts of orders and were therefore only used for the trade of infrequently traded stocks or the use was restricted to small orders. In 1991 the Frankfurt Stock Exchange introduced the *IBIS* trading system, where all major stocks could be traded.[69] But this trading system is not regarded as the official stock exchange, which still is a conventional exchange, it has initially been established as a system for inter-bank trading. With the reduction of the minimum order size to 100 shares in 1998 this trading platform now is also open to a wide public. The first official stock

[66] See COHEN ET AL. (1986, p. 49).

[67] The *SuperDOT 250*, implemented in November 1984 on the NYSE could handle a daily volume of 250 million shares. The capacity has been extended since then. Nowadays more than 1 billion shares can be traded on this system without problems. See SCHWARTZ (1988, p. 27).

[68] In 1982 a similar system was introduced on the Tokyo Stock Exchange, see SCHWARTZ (1988, p. 27), and 1986 on the Paris Stock Exchange, see SCHWARTZ (1993, p. 90).

[69] A new version with the name *XETRA* is used since November 1997. UNSER AND OEHLER (1998) provide a concise introduction into the features of this system.

exchange to introduce a fully automated trading system for all traded securities and order sizes, including an electronic clearing of trades, was the Swiss Exchange in Zürich (EBS) in 1996. In the mean time, more and more exchanges have introduced an electronic exchange, at least for a part of their trading activities.[70]

In dealer markets, also with a fully automated trading system, a minimum of inference is needed. The system can automatically route the orders to the market maker quoting the best price, who then only has to confirm the trade and enter his new prices into the system. In matching or batch systems a match maker is no longer needed, his duties can be taken over entirely by the computer system.

With the introduction of a fully automated trading system, a trading floor is no longer needed. All market participants only have to be connected to the computer system of the exchange. Their locations have no importance, they can trade from any place around the globe. If market participants gain access from another country this is called *remote access* and enables twenty-four hour trading on the same exchange if the market participants are located in different time zones.[71]

Computerization *per se* does not change the market structure or forces to do so.[72] In many cases, however, the introduction of a new computer system is used to establish a new market structure to increase the efficiency of trading. In many cases order book systems are introduced, replacing or complementing market makers, like on the London Stock Exchange.

Nevertheless, computerization has had a great impact on trading: orders can be executed more accurate to the rules, as errors are reduced, further they can be executed faster by a computer system, trading costs are lowered and trade information (volume and prices) can be transmitted faster. Investors are also able to react more

[70] Especially the futures exchanges e.g. EUREX or LIFFE have introduced electronic trading.
[71] The GLOBEX system of the CME, introduced in 1992, was the first, and by now the only system that enables a twenty-four hour trading on the same exchange.
[72] As has been pointed out above, only exchanges with crowd trading are not able to preserve their market form, as this form needs the physical presence of the brokers on the trading floor to negotiate the price.

quickly to changes in the market, as access to real-time data has become affordable to a wide public[73] and investors can submit their orders electronically to their broker, e.g. by using the internet.

Before the computerization of exchanges, a broker had to find the best price for an order, what in dealer markets with many competing market makers was a difficult task, because the quotes may change after every trade. In matching markets it was important to transmit the orders as soon as possible to the exchange before the prices changed too much. With a computerized exchange the best price is found by the computer system, the broker only has to transmit the order he receives - more and more electronically - from his client. His service is reduced to a pure transmission of the order. Due to this computerization and the increased global competition of brokers, broker fees have decreased significantly in the last years, reducing the costs to investors.[74]

As the broker no longer plays an active role, the submission of an order via a broker, who immediately passes this order without any interferences to the exchange, has the same effect as if the order would be directly submitted to the exchange, i.e. as if the investors were directly interacting.[75] Therefore the process of computerization is also called the *disintermediation* of financial markets.

2.7 Recent developments

Stock and futures exchanges are currently challenged by a number of changes taking place. First of all the computerization has increased the *competition* between

[73] Several brokers display real-time data for their customers on the internet for free. The costs of data providers also have decreased substantially in recent year due to competition from the internet.

[74] For a trade of about USD 10,000 several brokers offer fees of less than USD 10 when the order is placed using the internet, compared to fees of more than USD 100 not long ago and still applied by conventional brokers.

[75] The only reason, besides regulatory restrictions, that brokers are still used as intermediaires, is to reduce the counterpart risk of a trade as brokers guarantee to fulfill the trades of their customers.

Partners	Year	Features
DTB, SOFFEX	1998	common trading platform EUREX
NASDAQ, AMEX	1998	merger, separate markets
Paris, Zürich	1999	reciprocal access of members
EUREX, CBoT	1999	reciprocal access of members
LIFFE, CME	1999	reciprocal access of members
Paris, Lissabon	1999	reciprocal access of members
London, Frankfurt, Paris, Milano, Amsterdam, Brussels, Zürich	1999 (*)	common listings of stocks planned
NYSE, NASDAQ	2000 (*)	talks about merger
Paris, Brussels, Amsterdam	2000	merged to Euronext
Frankfurt, London	2000	merger to iX announced, later withdrawn due to a hostile takeover bid for the LSE from OM Gruppen
iX, NASDAQ, Madrid, Milano	2000 (*)	talks about merger

Those alliances under discussion are marked by an asterisk (*).

Table 2.2: Overview of alliances between major stock and futures exchanges

exchanges. Due to the possibility of remote access in computerized markets, the physical location of an exchange close to the most important financial institutions has become less important. This development increased competition globally between exchanges for the listing of securities as well as for order flow.

In this competition trading costs are a very important factor. Therefore exchanges have to improve their market structure to reduce trading costs for investors. Increased fixed costs for developing, improving and maintaining the computer systems of an exchange with fast growing trading volumes lead to numerous *alliances* between exchanges, seeking economies of scale. The number of these alliances, cooperations and mergers are advancing very fast. Table 2.2 lists some of the most important alliances, which in most cases guarantee reciprocal access to the partners for all members of one exchange. In some cases a common listing of securities, a common trading platform or even a merger is planned. However, with exception of the EUREX trading platform[76] and the NASDAQ/AMEX-merger none of the listed

[76] A brief description of the EUREX trading system is given in SCHILLER AND MAREK (2000).

ECN	March 1999	January 2000	March 2000
Instinet	18.0	14.4	14.1
Island	7.5	7.7	6.4
REDI-Book	.9	1.3	1.8
Brut	.7	.8	1.0
Archipelago	.7	.5	1.0
Tradebook	.8	1.3	1.5
Strike	.1	.1	.1
Nextrade	$< .1$	$< .1$	$< .1$
Attain	.1	$< .1$	$< .1$
Total	**28.8**	**26.2**	**26.1**

All numbers in %

Table 2.3: Market share in Dollar volume of ECNs on the NASDAQ

alliances is currently in operation. Most alliances are still in the state of planning how to realize the announced measures.

Typically stock exchanges are organized as non-profit organizations supported by its members, in general the financial institutions having direct access to the market. Growing investments into the computerization made it more and more difficult for exchanges to raise the necessary capital from its members. Therefore several exchanges discuss the transformation into *for-profit corporations* which would give them more flexibility in raising capital.[77]

This development is further accelerated by the increasing number of *Electronic Communications Networks* (ECNs), which are private-owned trading platforms offering similar services as an exchange. These ECNs, which are mostly financed and operated by large financial institutions, have gained a substantial market share in trading NASDAQ securities.[78] Table 2.3 shows that ECNs have a gained a market share of more than a quarter in dollar trading volume of securities listed on the NASDAQ,

[77] Such plans are under discussion at the Frankfurt Stock Exchange, the NYSE, the NASDAQ, and the CBoT. The members of the LSE and the CME already have approved such plans on March 15, 2000 and June 6, 2000, respectively.

[78] The market share for securities listed on the NYSE can be neglected as regulation does generally not allow members of the NYSE to trade securities listed on the NYSE off exchange. Up to now in Europe only a single ECN concentrating on securities listed on the LSE, Tradepoint, exists, which has a negligible market share of less than 1% in trading volume and gained the official status as an exchange. Other ECNs, like Jiway are planned.

but only two ECNs, Instinet and Island, contribute substantially to this result. The other ECNs are catching up significantly in the last months and it is expected that ECNs can increase their market share further.[79]

A final development is the extension of *trading hours*. Especially institutional investors are demanding longer trading hours, which enable them to react faster on news they receive. Since May 15, 2000 the Milano Stock Exchange trades 65 large stocks until 8.30 pm and the Frankfurt Stock Exchange extended their trading hours for all listed stocks until 8 pm on June 2, 2000.[80]

[79] GOMBER (2000) gives an overview of the requirements for a successful electronic trading system based on economic considerations.

[80] The traditional trading hours on the Frankfurt Stock Exchange have been from 10.30 am to 1.30 pm, currently most European markets operate from 9 am to 5 pm, the NYSE operates from 9.30 am to 4 pm. Extensions to 8 or 10 pm are planned by most leading stock exchanges in the near future.

Chapter 3

The NASDAQ Stock Market

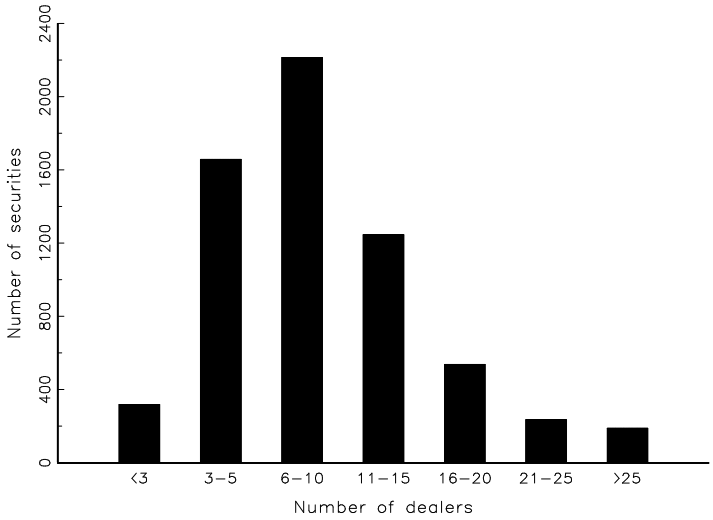
In chapter 2 possible market structures have been described, this chapter will give an overview of the structure of a specific stock market, the *NASDAQ Stock Market*. Although the *New York Stock Exchange* (*NYSE*) is the dominant exchange not only of the United States but of the entire world,[1] it faces fierce competition, especially from the NASDAQ Stock Market.[2]

Increased public attention has been paid in recent years to this market as many companies operating in fast growing sectors like information technology, biotechnology or telecommunications are listed on the NASDAQ. While the number of companies listed on the NYSE grew only slowly in the past years, the NASDAQ was able to attract a much larger number of companies. An increasing number of non-US companies consider to be listed on the NASDAQ rather than on the NYSE.[3] This increased importance of the NASDAQ also resulted in widened attention of NASDAQ trading rules in the academic literature, especially after the findings of Christie and Schultz (1994) on implicit collusion among NASDAQ market makers.

[1] When mentioning *"Wall Street"* this mostly is referred to the NYSE, which is located at 11 Wall Street in New York, but it is also used as a synonym for any stock exchange in the United States.

[2] The importance can be seen from the wide dispersion of computers displaying NASDAQ quotes. In 1996 300,613 such computers were operated within the United States and 37,846 in other countries. The countries with the largest number of these computers were Canada (16007 computers), Switzerland (6731) and the United Kingdom (5984), see The NASDAQ Stock Market, Inc. (1997, p. 32).

[3] 418 foreign companies have been listed on the NASDAQ in 1997 compared to 343 on the NYSE.



Data: NASDAQ Factbook 1997

Figure 3.1: Distribution of the number of market makers for NASDAQ securities in 1996

The NASDAQ is a dealer market with many market makers competing for the order flow of a specific security. For most securities there are between 3 and 15 market makers, with an average of about 10 market makers per security. Every market maker makes the market for nearly 100 securities on average. Figure 3.1 shows the distribution of the number of market makers per security.

3.1 History of the NASDAQ Stock Market[4]

With the Securities Exchange Act of 1934 every registered securities exchange was allowed to issue their own rules for admission to the market, listing of securities and trading within the framework of the Act. Exchanges were established as self-regulatory organizations. The Act did not encompass the trading taking place in securities not listed on a securities exchange, i.e. traded on the *over-the-counter market* (*OTC market*).

[4] This section follows SMITH ET AL. (1998, pp. 3-14) unless otherwise stated.

The Maloney Act of 1938 amended the Securities Exchange Act to provide also a framework for OTC transactions. It allowed the establishment of national securities associations that would issue guidelines for OTC trading and serve as *self-regulatory organizations* for their members. The only such association ever founded was the *National Association of Securities Dealers, Inc.* (*NASD*) in 1939. As the Act allowed members of such associations to discriminate against non-members in trading, the number of members increased from a portion of 22% of all firms engaged in securities trading in 1939 to 83% in 1982. Those firms not being members of the NASD where only regulated by the Securities and Exchange Act, while members were also subject to the rules of the association. In order to simplify and standardize supervision, in 1983 all firms engaged in the OTC markets were forced to join a national securities association. As the NASD was the only such association, nearly all companies trading securities became members, with the exception of those few only trading on a registered exchange, as the NYSE, Amex or one of the five regional exchanges.[5]

A characteristic of OTC markets is that market participants are not centralized on a trading floor like the NYSE, but that they are dispersed all over the country or even the entire world. A transaction before the start of computerization typically occurred as follows: an investor submitted an order to a broker, the broker then tried to find out the market maker quoting the best price.[6] As a medium of communication in most cases the telephone had been used, the broker had to phone the market makers and ask for their quotes.[7] The broker traded with the market maker quoting the best price on his own account at the stated price. The customers had been charged a mark-up on this price to cover the expenses of the market maker.

[5] The remaining regional exchanges in the United States are: Boston Stock Exchange, Chicago Stock Exchange, Cincinnati Stock Exchange, Pacific Stock Exchange, and Philadelphia Stock Exchange.

[6] OTC markets are typically organized as dealer markets. By having established a market maker it is easy to find a counterpart that otherwise would hardly be found as a result of decentralization.

[7] Often it happened that market makers and brokers were identical and the search process therefore was simplified, but it remained difficult to determine whether there existed a more favorite price from another market maker.

The prices of the market makers were not published real-time due to the lack of any technologies enabling this, quotes (typically average or closing quotes) were published the following day in printed bulletins only available to brokers and market makers.

Such a trading mechanism has the disadvantage that finding the best quoted price is very difficult, time intensive and may not yet be found, such that transactions occurred at less favorable prices. Transaction costs of trading will be high due to the lack of transparency in such a market. Furthermore the reaction of investors to new information was difficult as the informativeness of prices has been low. These inefficiencies of OTC markets made them not very attractive for investors and companies, hence for a long time they were no meaningful competitors to registered exchanges. Often for small companies it was the only possibility to raise new equity by being traded on OTC markets, most of them applied to be listed on an exchange when they fulfilled the listing requirements.

Improvements in telecommunications and computer technologies during the 1960's enabled quotes to be disseminated faster. In 1966 the NASD began to consider an automated quotation system that would allow real-time quotes to be displayed on screens connected with a central computer. The market makers would have been able to enter their quotes and the best quotes were to be displayed on screens, mentioning the market maker quoting it. With this information a broker could directly address the market maker quoting the best price and was no longer forced to call all market makers in order to get this information. Under the name *National Association of Securities Dealers Automated Quotation System* (*NASDAQ*) this system was put into operation on February 8, 1971 by linking about 500 market makers, a large number of brokers and even more interested parties, like investment consultants, to a central computer. Three different levels of service had been established: The *level 1 service* allowed to follow the market by observing the best available quotes. This service was designed for investment consultants and the public. With a *level 2 service*, available for institutional investors and brokers, it was possible to observe

not only the best quotes, but all quotes of the market makers. The names of the market makers quoting the prices were also displayed. With *level 3 service* market makers were able to enter their quotes.[8]

To be listed on the NASDAQ, companies had to meet minimal requirements with respect to size and corporate governance, otherwise they were not included into this new system.[9] Trading securities listed on the NASDAQ changed significantly. The determination of the best available price had become much easier and the transparency of the market increased. Also the settlement between brokers and their customers changed, brokers did no longer charge their customers a mark-up on the price they received from the market maker, but charged the same price and instead used commission fees to cover their costs.

Initially the NASDAQ had been designed only to disseminate information on quotes, information on trades having occurred could not be obtained. Improved computer technologies, however, allowed to provide these information for 40 of the most active securities in 1982. In due time more securities that met requirements more restrictive than being listed on the NASDAQ were incorporated into this new service, called *NASDAQ National Market System* (*NASDAQ/NMS*).[10] In 1983 682 securities were listed on the NMS, 2587 in 1990 and 4371 in 1996. Those securities not included into the NMS are mostly small and infrequently traded. These *regular NASDAQ*[11] securities were traded only with quote information until 1992, when also information on trades has been added.

By launching the *Small Order Execution System* (*SOES*) in 1984 the NASDAQ became a trading platform rather than only a tool for information dissemination of quotes and trades. The SOES enabled orders to be automatically routed to the market maker quoting the best price. The market maker only had to confirm the execution of orders by pushing a button, a confirmation of order execution is

[8] See SCHWARTZ (1993, p. 53).
[9] These requirements are stated in chapter 3.3.
[10] In 1993 renamed into *NASDAQ National Market* (*NNM*).
[11] In 1993 this tier of the NASDAQ has been renamed into *NASDAQ¿SmallCap Market.*

sent to the broker electronically without further personal interference. Originally participation in the SOES was voluntary for market makers, the use was restricted to NMS securities and order sizes of 500 shares or below. In 1985 all securities were included and the maximum order size for NMS securities raised to 1000 shares. During the crash of 1987 market makers were difficult to reach by phone and many orders could not executed within an acceptable time. In reaction to this experience, participation in the SOES for NMS securities became mandatory for all market makers in 1988.

The same experience during the crash of 1987 led to the development of the *Order Confirmation Transaction Service* (*OCT*) in 1988,[12] where orders could be submitted electronically to a specific market maker instead of using the phone. By pushing a button to confirm the execution of the order, this system enables faster execution of trades, hence larger trading volumes can be processed than by using the phone only.

Also in 1988 the *Advanced Computerized Execution System* (*ACES*) has been introduced. Participation in this system is voluntary for market makers and brokers. It allows orders to be automatically routed to the best participating market maker and the execution is again confirmed only by pushing a button. Unlike in the SOES, the order size is not restricted by the system. Every market maker participating in this system has to negotiate with one or more brokers up to which order size he is willing to execute the orders at the stated prices.[13] He can negotiate different order sizes with different brokers and for different securities. A negotiation with all brokers is not necessary. Between large brokers and market makers similar private systems exist, especially in cases where brokers and market makers are employed by the same financial institution.

Since these developments the systems have continually been improved to be easier

[12] An improved system has been introduced in 1990 under the name *SelectNet*.

[13] As will be presented in section 3.5 the trading rules require the quotes to be valid for a minimum order size. This system enables market makers and brokers to negotiate a higher order size bilaterally.

Year	Event
1939	Foundation of NASD
1971	NASDAQ starts operation as a quote dissemination system
1982	Introduction of a two tier market with dissemination of trade information for the National Market
1984	SOES launched with mandatory participation
1988	OCT introduced ACES introduced SOES becomes mandatory for National Market securities
1992	Dissemination of trade information for the SmallCap Market

Table 3.1: Main historical events of the NASDAQ

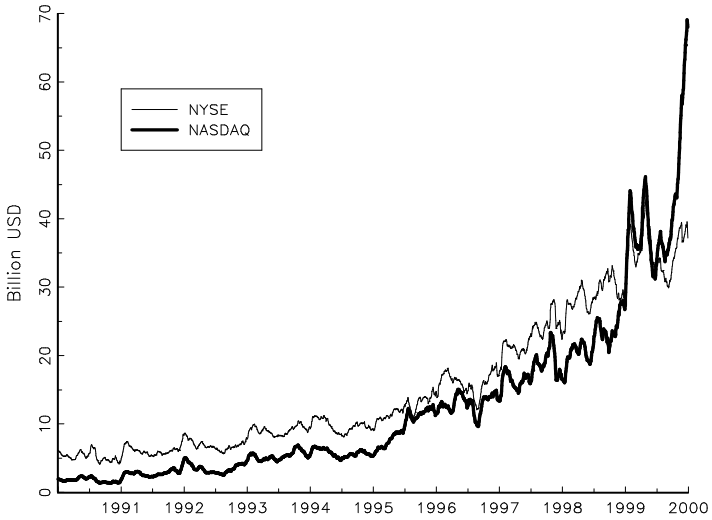
to handle and to be able to conduct an increasing number of trades. A new system called *NASDAQ Order Delivery and Execution System* (*NODES*) is currently awaiting approval by the Securities and Exchange Commission. It shall replace and improve the current SOES and SelectNet.[14]

The progress in trading transparency and increased standards in regulation has made the former OTC market comparable to a securities exchange, consequently in most minds it is regarded as an exchange. Table 3.1 summarizes the main events in the history of the NASDAQ. The NASDAQ is today widely accepted by investors and companies as a market comparable to the NYSE. It has become the largest market of the world in dollar and share trading volume ahead of the NYSE and the second largest in market capitalization, just behind the NYSE. In recent years it has significantly catched up with the NYSE and in many respects surpassed it.

Most recently the NASDAQ Composite Index outperformed the Dow Jones Industrial Average Index, which mostly consists of stock listed on the NYSE.[15] In combination with more attention being paid to internet and biotechnology stocks, which are mostly listed on the NASDAQ, the market received more and more interest from the general public. Figures 3.2 to 3.4 illustrate these recent developments.

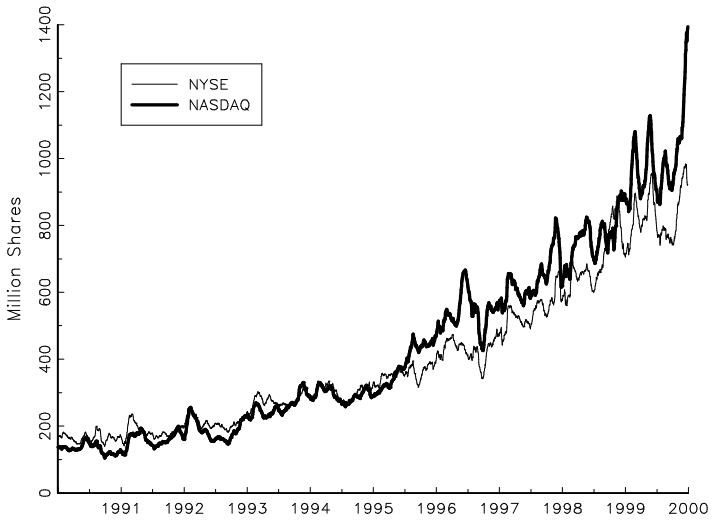
[14] See *Research Matters* 1(2), 1998, p. 4, published by the NASD Economic Research Department.

[15] Only in late 1999 the Dow Jones Index included large companies listed on the NASDAQ, like Microsoft or Cisco Systems.



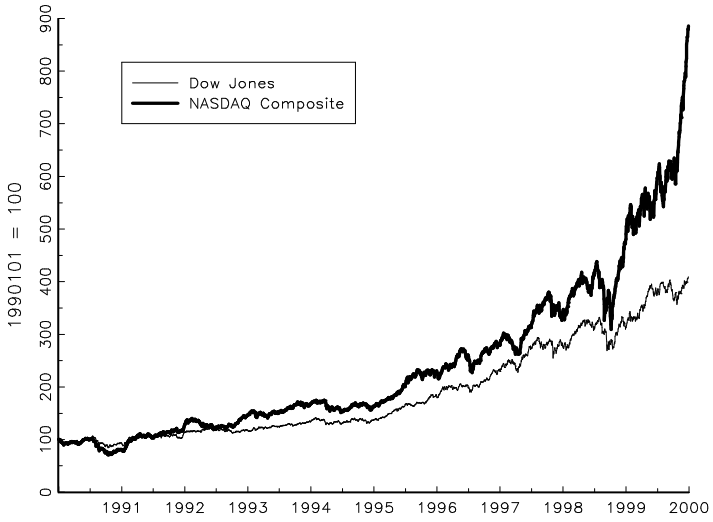
Data: NYSE and NASDAQ

Figure 3.2: Daily US-Dollar trading volume on the NYSE and the NASDAQ (20 day moving average)



Data: NYSE and NASDAQ

Figure 3.3: Daily share trading volume on the NYSE and the NASDAQ (20 day moving average)



Data: Datastream

Figure 3.4: Development of the Dow Jones Industrial Average and NASDAQ Composite Index

Many large companies, although fulfilling the requirements to be listed on the NYSE, such as Microsoft or Intel, remain to be listed on the NASDAQ and also many foreign companies decide to be listed on the NASDAQ rather than on the NYSE. This gives evidence that the NYSE and NASDAQ have become equal competitors. Recent improvements in the transparency of the markets are attributed to the competition for trading volume and the listing of companies.

3.2 The organization of the NASD[16]

The *NASD* is a self-regulatory organization supported by its members, brokers and market makers trading on OTC markets. The NASD itself only operates departments that serve the organization as a whole, such as economic research, human resources or finance. All operations are conducted by three subsidiaries, the NAS-

[16] This section is based on SMITH ET AL. (1998, pp. 14-20).

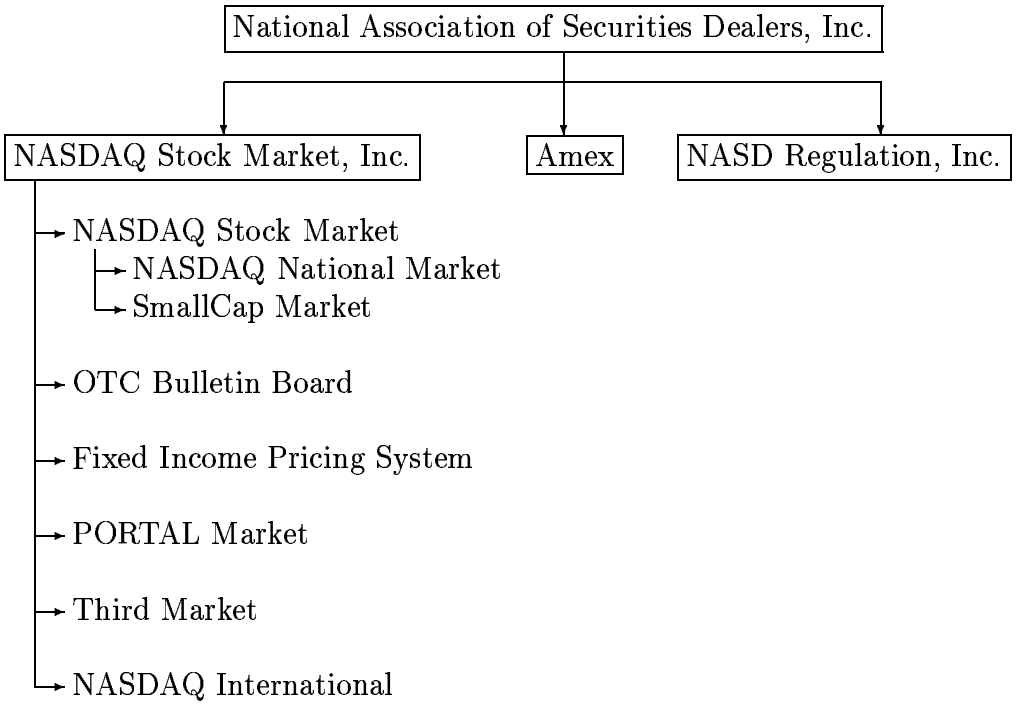


Figure 3.5: Organization of the NASD

DAQ Stock Market, Inc.[17], the American Stock Exchange (Amex) and the NASD Regulation, Inc. Figure 3.5 shows the organizational structure of the NASD.

The *NASDAQ Stock Market, Inc.* operates the different OTC markets. It develops and maintains the computer and telecommunications networks used for market operations and develops new trading systems. It further promotes the listing of securities and investments into NASDAQ securities, e.g. through sponsoring or seminars.

The NASDAQ Stock Market, Inc. runs several OTC markets, most prominent is the *NASDAQ Stock Market* with its two tiers, the NNM and the SmallCap Market. These markets are mostly referred to as NASDAQ, a convention that will also be used in this work. The other markets are of less importance and therefore receive only limited attention. The *OTC Bulletin Board* (*OTCBB*) is a pure quotation system for securities not listed on the NASDAQ or any exchange. No trade information is displayed and no trades can be conducted or initiated through this system. The

[17] The NASD currently considers plans for a going public of the NASDAQ Stock Market, Inc.

Fixed Income Pricing System (*FIPS*) is a quotation and trade information system for about 50 of the most actively traded high yield corporate bonds (rated BB+ or lower by Standard & Poor's). The *Private Offerings, Resales and Trading Through Automated Linkages* (*PORTAL*) *Market* allows private placements of securities to be better allocated by publishing information on prices and the securities themselves. This market is restricted to institutional investors with an investment of at least USD 100 Mio. in security markets. It also provides a platform for trading those securities, but this possibility is rarely used. In the *Third Market* securities listed on a registered exchange can be traded off the exchange by using certain facilities of the NASDAQ and applying similar rules. The attempt to offer trading in NNM securities and selected foreign securities at European trading hours at *NASDAQ International* operating in exactly the same way as the NASDAQ Stock Market using its computer facilities, has not generated much interest thus far. Currently the NASDAQ is planning to expand their activity to Japan, Canada and Europe through building up new trading platforms and seeking cooperations with established stock exchanges and private trading platforms.

The *American Stock Exchange* (*Amex*) has become a subsidiary of the NASD since their merger has come into effect on October 30, 1998. It is a registered exchange and is operated independently of the other markets.

The *NASD Regulation, Inc.* (*NASDR*), founded in 1996, has overtaken all regulatory affairs that have formerly been conducted directly by the NASD.[18] Besides defining rules for trading on NASDAQ markets, it also supervises the compliance to these rules and has the authority to sanction violations. If laws have been violated it informs the legal authorities and cooperates in investigations. The NASDR further administers written tests to qualify securities professionals and registers them.

[18] As SCHULTZ (2000) points out, it was the Christie-Schultz debate that forced the NASD to give its regulatory division more autonomy by founding a separate subsidiary.

3.3 Listing requirements

To be listed on the NASDAQ Stock Market, a company has to meet certain criteria regarding corporate governance, public disclosure, and size. The aim of these criteria is to ensure a minimum of investor protection and to enable an orderly trading process.

According to Rule 4310 all companies have to meet the following qualitative criteria:[19]

- at least two independent directors on the board,
- an independent internal audit committee,
- an independent public accountant auditing the company,
- an annual report to be distributed to all shareholders,
- an annual meeting of shareholders,
- important corporate actions have to be approved by the shareholders,
- a quorum of at least $^1/_3$ of all outstanding shares for all decisions of the shareholders,
- prompt disclosure of information through the media that affect the value of the shares.

Additionally to these qualitative criteria, companies have to meet certain quantitative standards. To be listed on the SmallCap Market Rule 4310 requires the companies to meet these minimal criteria:

- net tangible assets: USD 4 Mio.,

[19] These criteria can be adapted for foreign companies to meet the regulatory framework in their country of residence.

- market capitalization: USD 50 Mio.,
  or net income: USD 750,000,
  or operating history: 1 year,
- number of publicly held shares: 1 Mio.,
- round lot shareholders[20]: 300,
- minimum bid price: USD 4,
- number of market makers for the security: 3.

To be listed on the NNM Rule 4420 requires that one of the following three standards has to be met:

- Standard 1:
  - net tangible assets: USD 6 Mio.,
  - pretax income: USD 1 Mio.,
  - number of publicly held shares: 1.1 Mio.,
  - market value of publicly held shares: USD 8 Mio.,
  - round lot shareholders: 400,
  - minimum bid price: USD 5,
  - number of market makers: 3.
- Standard 2:
  - net tangible assets: USD 18 Mio.,
  - operating history: 2 years,
  - number of publicly held shares: 1.1 Mio.,
  - market value of publicly held shares: USD 18 Mio.,

[20] Shareholders holding at least shares of one trading lot, in most cases 500 shares.

  - round lot shareholders: 400,
  - minimum bid price: USD 5,
  - number of market makers: 3.

- Standard 3:
  - market capitalization: USD 75 Mio.,
    or total assets: USD 75 Mio. and total revenue: USD 75 Mio.,
  - number of publicly held shares: 1.1 Mio.,
  - market value of publicly held shares: USD 20 Mio.,
  - round lot shareholders: 400,
  - minimum bid price: USD 5,
  - number of market makers: 4.

Once a company is listed, it can fall short of these quantitative criteria, but not of the qualitative criteria. In order to maintain the listing similar, less restrictive criteria have to be met according to Rules 4310 and 4450.

Additional to these criteria, companies have to pay an entry fee for being listed and an annual fee to maintain the listing. These fees depend on the market on which the company is listed and its size. The entry fee for a listing on the SmallCap Market is between USD 5,000 and USD 10,000, for a listing on the NNM between USD 5,000 and USD 50,000. The annual fees are USD 4,000 for the first security of a company listed on the SmallCap Market and USD 1,000 for each additional security. In the NNM this fee varies between USD 5,250 and USD 20,000.[21]

3.4 Registration as broker and market maker

There exist two prerequisites to register as broker or market maker. The first concerns the capital requirements to ensure those market participants to conduct their

[21] See Rules 4510 and 4520.

duties without facing the threat of bankruptcy. These prerequisites are regulated by the *Securities and Exchange Commission* (*SEC*). The other prerequisites refer to their qualifications and are regulated by the NASD.

A *broker* has to maintain a net capital[22] of at least USD 100,000. A *market maker* needs a net capital of USD 2,500 for each security he makes the market in.[23] Additionally, a minimum of USD 100,000 and a maximum of USD 1 Mio. applies. In most cases brokers and market makers are companies rather than individuals, whose business is conducted by employees. In this case the company as a whole has to fulfill these requirements, it has not to be fulfilled for every single employee acting as market maker or broker.[24]

The Securities Exchange Act requires every broker and market maker acting on OTC markets, like the NASDAQ Stock Market, to be member of a national securities association, hence they have to be member of the NASD. Not only the brokerage companies and companies acting as market makers have to be registered, but the by-laws of the NASD require every employee of those companies who is involved in brokerage or market making activities to become a member. While companies are registered with approval of the SEC, their personnel has to prove their qualifications to become members. Without being registered as member, no individual is allowed to conduct businesses related to brokerage or market making.

The by-laws of the NASD require members to have an appropriate qualification to conduct the business they are assigned to.[25] According to Rules 1021, 1031 and 1041 these qualifications have to be proved by passing a qualification examination conducted by the NASDR. These examinations are designed to explore the qualifications for a specific duty. If the duties of a person change, it can be necessary to pass another examination for his new duties. After having passed the examination the person is registered as member and allowed to conduct the business he has been

[22] Net capital is the net worth adjusted for unrealized profits and losses, subordinated loans, etc.
[23] For securities with a market value of less than USD 5 per share the requirement is USD 1,000.
[24] See SEC Rule 15c3-1.
[25] See Article III, Section 2 of the by-laws of the NASD.

assigned to. Furthermore, Rule 1120 requires registered members to follow certain continuing education requirements.

When registered as broker, one is free to act as broker for all securities listed on the NASDAQ Stock Market. Being registered as market maker also allows to become market maker in every listed security, it is only necessary to register for the securities one wants to make a market in. Market making can begin the next trading day.[26] With registration as market maker for a specific security it is the obligation to quote always a price at which one is willing to buy and sell the security.[27]

To withdraw the registration as market maker for a specific security follows a similar process. A request to withdraw the registration for this security becomes effective the next trading day. The only restriction faced upon withdrawal is that it is not allowed to register again as market maker for the same security during the next 20 trading days.[28] This free market entry and exit in most cases leads to more than two market makers being registered for a security. Furthermore, it is the aim to prevent registered market makers from making extraordinary profits through their activities by imposing the threat of new market makers entering.

3.5 Trading rules

A market maker registered for a security has the obligation to quote prices both for buying and selling the security from the public during the trading hours from 9.30 am to 4 pm.[29] The monthly average spread a market maker quotes for a security must not exceed 150% of the average spread of all market makers for this security. Rule 2440 and Interpretation IM-2440 require the market makers not to charge a too large spread. As a guideline a maximum spread of 5% is mentioned in this rule, depending on the circumstances, e.g. market conditions and characteristics of the

[26] See Rule 4611.
[27] See Rule 4613. Section 3.5 describes the trading rules in more detail.
[28] See Rule 4620.
[29] See Rules 4613 and 4617. Extending trading hours to 8 or 10 pm is currently considered by the NASDAQ.

security. Larger spreads can be justifiable, but also a spread of 5% may be viewed as too large by supervisors of the NASDR, forcing market makers to reduce their spread. However, no fixed rule can be applied to determine the maximum spread, it is subject to interpretations by the NASDR.

The quotes are further restricted by tick sizes, the increments have to be multiples of USD $^1/_{32}$ for securities with bid prices below USD 10 and USD $^1/_{16}$ for those above.[30] Quotes have to be firm, i.e. upon request the market maker has to trade at least at the stated prices, but he is free to choose a more favorable price for the transaction.[31] The obligation of a firm quote is only waived for a short period of time to enable the market maker an update of his quotes after having executed an order.[32]

Furthermore, quotes have to be valid at least for a normal trading size, a lot of 100 shares. The number of shares a market maker is willing to trade at the quotes are displayed on the screen next to their quote. Rule 4613 requires the minimum trade sizes for which the quotes have to be valid to be larger than 100 shares under certain conditions. For securities listed on the SmallCap Market the minimum trade sizes is 500 shares if the average daily non-block volume[33] exceeds 1000 shares or the bid price is below USD 10.

For securities listed on the NNM these limits are:

- 1000 shares if
  - the average daily non-block volume is above 3000 shares,
  - the bid price is below USD 100, and

[30] See SMITH ET AL. (1998, p. 28). Changing to a decimal system is considered for late 2000. Through 1997 the tick size has been USD $^1/_8$ for securities with a bid price above USD 10 and USD $^1/_{16}$ for securities with a bid price of USD 10 or below.

[31] When choosing a more favorable price he is not restricted to the tick sizes in determining the price he charges. The tick sizes only apply to quotes, not to transaction prices. More favorable prices than those quoted are frequently observed as a result of preferencing arrangements, which are described below.

[32] See Rule 3320 and interpretation IM-3320.

[33] For the definition of block trades see below.

  - there are at least 3 market makers for this security.

- 500 shares if
  - the average daily non-block volume is above 1000 shares,
  - the bid price is below USD 150, and
  - there are at least 2 market makers for this security.

- 200 shares if
  - the average daily non-block volume is below 1000 shares,
  - the bid price is below USD 250, and
  - there are at least 2 market makers for this security.

Orders that are larger than the market makers are willing to accept, can be broken into parts of at least a lot and be executed like several smaller orders. This may result in different prices applied for each part and the parts may be executed by different market makers. Limit Orders may also be executed in parts of at least a lot with offsetting orders. To avoid partial execution the order has to be specially marked as a *All-or-None order* by the investor.

Trades of 10,000 shares and above are called *block trades*. Such trades are subject to special treatment. They can either be traded through market or limit orders as a whole or be broken into several smaller orders within the normal trading procedure. Investors face the risk of influencing the price significantly in an unfavorable way through the placement of such an order. For this reason such orders are usually traded in a special market (*upstairs market*) for separate negotiation with other block trades.[34]

An order arriving on the market has to be executed at the best available price (*price priority*), i.e. the market maker quoting the most favorable price has to execute

[34] See SCHWARTZ (1993, pp. 41ff.).

the order at the stated or an even more favorable price.[35] If several market makers quote the same price, the market maker executing the order can be chosen without restrictions by the brokers. Preferencing arrangements, as described below, are applied in most of these cases to determine the routing of the order flow.

Price priority and interpretation IM-2110-2 of Rule 2110 give the guidelines for handling *limit orders*. Limit orders can be accepted by market makers, but they do not have to be. By accepting a limit order, the market maker has to follow the established rules. He must not trade ahead of a limit order he has received, i.e. is not allowed to execute an offsetting order on his own account at the same or a less favorable price than the limit has been set (*public before dealer*). A market maker can immediately execute limit orders on his own account or can route them to other market makers. To enhance the transparency of the market, SEC Rule 11Ac1-4 requires unexecuted limit orders to be displayed in the quotes of a market maker. If the limit order has the best available price, also its size has to be displayed. The obligation is only waived for orders below 100 and above 10,000 shares and if it must not be partially executed (All-or-None orders).[36]

The aim of these rules on the handling of limit orders is to guarantee a maximum of transparency and enhance competition further by allowing limit orders directly to compete with the quotes of market makers.[37]

As has been stated above, *preferencing* in most cases determines the market maker who executes an incoming order in the case where several market makers quote the best price. With preferencing a broker routes his entire order flow to one specific

[35] See Rule 2320.

[36] See SMITH ET AL. (1998, p. 24).

[37] Through 1994 limit orders were interpreted to be offers of investors to trade with a market maker at the stated price. Hence market makers could trade with other investors on their own accounts at less favorable prices, i.e. higher ask and lower bid prices. Limit orders were only executed against quotes of the market makers and not orders from other investors, consequently they also have not been published. From 1994 onwards, market makers were still allowed to trade ahead of limit orders, but only if they quoted the same or a more favorable price, limit orders had not to be published. In 1997 the current regulation has been introduced. Allegations of market makers colluding on wide spreads in 1994 lead to these changes in the handing of limit orders to enhance competition. We will address the competition between market makers and limit orders in chapter 6.1 in more detail.

market maker, provided he quotes the best available price and price priority can be applied.[38] The two main reasons for such a behavior are either internalization or payment-for-order-flow.

In many cases companies act both as broker and market maker. In this case vertical integration results in preferencing, what is also called *internalization*. To maximize profits, the brokerage department has to route all orders to the own market makers if they quote the best available prices.

A broker may also be willing to route his order flow to a specific market maker because he receives a payment from him (*payment-for-order-flow*). This payment can either be in form of cash, or the market maker charges a more favorable price to the broker than his quote. The broker can either receive this difference to the quoted price by charging his customer the quoted price or he can forward the whole or a part of this surplus to his customer and gain a competitive advantage over other brokers, either by charging a more favorable price or by reducing his commission fees.[39] Other forms of payments can also include various services, e.g. research reports on companies or conducting the clearing process. Payments typically have a value between USD .01 and USD .02 for each share.

Those market makers and brokers participating in ACES or similar private arrangements are very likely in preferencing arrangements. Preferencing adds another source of competition between market makers, besides price competition they also compete for trading volume.

[38] In many cases preferencing arrangements require that the entire order flow is routed to a specific market maker, regardless of his current quotes. To fulfill the requirement of price priority for investors, the market makers in turn guarantee to charge only the best available price, i.e. if necessary they improve their quotes. Such arrangements are also called *price matching arrangements*, they are discussed in more detail in chapters 5.4 and 6.4.

[39] See SMITH ET AL. (1998, pp. 28 ff.).

3.6 Summary

The NASDAQ has grown out of a telecommunications network for disseminating quote information to a network having all features of an exchange. This development has been made possible by improvements in computer and telecommunications technologies. Differences to registered computerized exchanges are only minor nowadays, so that the NASDAQ is mostly referred to as an exchange, although it misses this formal status and is "only" an OTC market.[40]

The NASDAQ is characterized by the competition of market makers for the best price and for trading volume. If admitted as market maker to the NASDAQ, there are virtually no entry barriers for market making in a specific security. This allows for hit-and-run competition, ensuring no extraordinary profits to be made by market makers, hence low spreads should be expected.[41] Together with rules ensuring high standards of transparency for the market, trading costs should be low. Other stock exchanges, like the NYSE, have to establish a much more complex set of rules to abandon the use of market power that arises as the result of high entry barriers or a lack of competition by granting monopolies of market making.

Listing requirements of the NASDAQ are much less restrictive compared to those of other stock exchanges, e.g. the NYSE, as it is designed for small companies. These small, in many cases highly innovative companies operating in fast growing industries, made the NASDAQ well known to be a market for high-tech stocks. High market transparency and a seemingly competitive trading environment induced many companies having grown to sizes that qualify for a listing on the NYSE to remain their listing on the NASDAQ. It has become the major competitor of the

[40] See SMITH ET AL. (1998, p. 51).

[41] This result can best be described with the theory of *contestable markets*. It states that the threat of new market participants entering in case of excess profits, forces market incumbents to charge competitive prices. The absence of entry barriers (legal restrictions, sunk costs) are the prerequisites for a contestable market. BAUMOL ET AL. (1988) provide a detailed overview of the theory of contestable markets. However, as we will show in chapter 5, implicit collusion between market makers will enable them, despite these competitive forces, to quote noncompetitive prices and receive excess profits.

NYSE for the listing of companies and the NASDAQ Composite Index is one of the most important indices of the world, having received increased attention in recent years.

Chapter 4

Market microstructure theory

The last two chapters showed that trading in actual asset markets follows certain rules and imposes costs that have to be taken into account in explaining the price formation. As these costs differ among market structures, we have to investigate the trading in a given market structure. For this reason asset pricing models assuming frictionless trading, like the CAPM, may be appropriate to determine the fundamental value of an asset, but actual prices can differ substantially from this fundamental value. These considerations gave rise to *market microstructure theories* developed in the last two decades.[1] Market microstructure theories are not interested in the fundamental value, which is assumed to be determined exogenously, but to model the *price formation* in asset markets with a given fundamental value.

After some pioneering work by DEMSETZ (1968) and BAGEHOT (1971) the first models developed by STOLL (1978) focused on dealer markets and inventory effects that influence the price setting of market makers. Starting with COPELAND AND GALAI (1983) for dealer markets and KYLE (1985) for auction markets, the attention changed from inventory effects to that of different information market participants have. Those models dominate the literature since this time, addressing important questions like informational efficiency of prices and market liquidity.

[1] The term *market microstructure* has first been introduced by GARMAN (1976). According to EASLEY AND O'HARA (1995, p. 357) it is

> "the study of the process and outcomes of exchanging assets under explicit trading rules."

A common feature of these models is that they neglect direct costs faced by intermediaries, so called *order processing costs*. Order processing costs include fees collected by the exchange, costs for maintaining the infrastructure of intermediaries, e.g. offices, computer facilities, or salary of employees. They only come into effect again in empirical investigations as will be shown in chapter 4.7. Neglecting order processing costs in the models allows to explore the effects arising from other cost components, namely inventory and adverse selection costs, in more detail.

This chapter wants to give an overview of the basic models in market microstructure theory. Due to the extensive literature, a complete survey of all models and empirical investigations cannot be the aim. The first section will analyze why different market forms exist and how they emerge. Thereafter we will investigate the price formation in auction markets before concentrating in the remaining sections on different aspects of dealer markets.

Throughout this chapter we will also provide detailed derivations of the results for the most important models as they are reported in the literature. In the only textbook on the market microstructure theory of financial markets, O'Hara (1995), these derivations are not presented and the original journal articles in most cases provide only a rough guideline how to proof the results. Furthermore, we simplified some models by making additional assumptions, either to concentrate on the main results of the model or to have a coherent framework with closely related models also presented here. By giving detailed proofs the reader not only gets a deeper understanding of the mathematics behind the results, but also of the necessity of assumptions and restrictions for showing the existence and properties of a solution. In cases where proofs are too long or repetitive to other models, we report only the results and their intuition in order to limit space.

4.1 The emergence of different market forms

Chapter 2.3 gave an overview of the different market forms and their advantages and disadvantages from the view of an investor. It was intuitively shown that investors may demand different market forms, depending on their preferences and the characteristics of the asset. These different market forms also have to be supplied, i.e. there have to be match makers and market makers. In this section we will derive conditions under which auction and dealer markets arise in an unregulated market.[2]

In both market forms, auction and dealer markets, we find specialized market participants, match makers and market makers. Contrary to investors they do not actively trade in the market but react to the demand of investors.

A match maker receives an order[3] from an investor and has to find an offsetting order on the market. When he has found such an order, he arranges the trade and determines the price. For these efforts he charges a fee to both investors. At the time investors submit their orders to the match maker, neither the price at which the trade is conducted nor the time the trade occurs is known. Investors face the risk that prices may change significantly in the mean time and trading decisions may not longer be optimal.

A market maker, on the other hand, publishes prices at which he is willing to buy and sell the asset on his own account from any investor. At the time the market maker publishes these prices, he does not know whether the next order arriving will be a buy or a sell order, but any order is immediately executed at the stated prices. Investors know the price they will receive upon submitting an order and also the time of trading. They face no risk after having submitted the order. This risk is now taken by the market maker, he does not know when another, offsetting order will arrive on the market and at which price he will be able to trade. This risk imposes

[2] VRIEND (1996) provides a more general model how intermediaires are established in an economy.

[3] Unless otherwise stated only market orders will be considered. We will consider effects arising from limit orders in more detail in chapter 4.5.

costs on the market maker for which he has to be compensated. Typically a market maker is not allowed to charge any fee to investors, therefore he has to incorporate his costs into the prices he charges. He will quote a higher price for selling the asset to an investor (*ask price*) and a lower price for buying the asset from an investor (*bid price*). As will be shown in sections 4.3 and 4.4 the ask price will always exceed the bid price, the difference between these two prices is called the *spread*.

The possibility for investors to trade at a fixed price without any delay is called *immediacy*. The willingness to trade with a market maker is the *demand* for immediacy and the willingness to act as market maker is the *supply* of immediacy.[4]

The following section will provide an equilibrium of immediacy before in chapter 4.1.2 a model is presented to determine the equilibrium market form.

4.1.1 Demand and supply of immediacy

GROSSMAN AND MILLER (1988) assume a market with N participants, a single risky asset and a riskless asset with a zero rate of return.[5] Of the N market participants $n < N$ face an exogenously given liquidity event, i.e. their holding of the risky asset is no longer optimal and they have to rebalance their portfolio by trading the risky against the riskless asset. What causes this liquidity event is of no importance here, we could easily assume the liquidity event to be caused by an exogenous need for changing the investment position, but also informational asymmetries could be incorporated. The other $N - n$ market participants face no need to change their portfolios. The n market participants facing a liquidity event are called *investors*.

We assume that there exist three time periods. At the beginning of period 1 the liquidity event occurs and at the end of this period a first round of trading takes place. At the beginning of period 2 new information on the fundamental value of the

[4] See GROSSMAN AND MILLER (1988, p. 618). Another form of supplying immediacy is by submitting limit orders as we will se in chapter 4.5.

[5] The assumption of a zero return can be justified by using the return of the riskless asset as a normalization, i.e. all returns have to be interpreted as excess returns.

asset is released and afterwards a second round of trading is conducted. In period 3 the fundamental value is fully revealed and the portfolios are liquidated at the fundamental value without further trading.

The price formation has to be such that after the second trading round all market participants hold their optimal portfolio, i.e. they have adjusted their portfolios according to the liquidity event and the new information released in period 2.

Investors can offset their liquidity event either by trading in period 1, in period 2 or in both periods. Assume without loss of generality that they trade either in period 1 or in period 2, but not in both periods.[6] When an investor decides to trade, depends on his exogenously given preferences and the costs in each period. Assume that n_1 investors trade in period 1 and n_2 investors in period 2, where $n_1 + n_2 = n$. Let the number of assets that have to be traded as a result of the liquidity event by investor i be denoted x_0^i, where with a positive amount the investor buys additional units of the asset and with a negative amount he sells the asset in exchange for the riskless asset. The sum of all orders has to equal zero because the overall holding of the asset does not change:

$$\sum_{i=1}^{n} x_0^i = 0. \tag{4.1}$$

But if some investors decide to trade in period 1 and others in period 2, the orders arriving in each period will not necessarily be balanced. In order to clear the market, there have to be some market participants facing no liquidity event that are willing to take offsetting positions enabling investors to trade. These market participants do not have to trade as they face no liquidity event, but they can take the offsetting positions in period 1 and offset the positions acquired in period 1 by trading again with the remaining investors in period 2. Those market participants that take voluntary offsetting positions in period 1 are called *market makers*. Let there be $M \leq N - n$ market makers.

[6] Without changing the argument we could include investors trading in both periods. In this case we can interpret them as behaving partially as investors and partially as market makers, which we introduce below.

With X_0^1 denoting the order imbalance in period 1 we define

$$X_0^1 \equiv \sum_{i=1}^{n_1} x_0^i. \tag{4.2}$$

Market makers have only to be concerned with the net order imbalance as all other orders can immediately be offset in the same period by matching the orders directly and only the excess demand of the period is offset by the market makers. Hence they are only concerned about trades in one direction, either they buy or sell the asset. As market makers have to quote fixed prices at which they are willing to trade, but only trades at one side occur, it has not to be distinguished between bid and ask prices. The price of the asset in period t, P_t, has to be interpreted as a bid or ask price, depending on the net order flow.

We assume that all market participants maximize their expected utility of final, i.e. period 3, wealth.[7] Denote the wealth of investor i in period t by W_t^i, the amount of the riskless asset he holds by B_t^i, and the units of risky assets held by q_t^i. Let further W_0^i be the initial wealth of an investor excluding the wealth to be reallocated as a consequence of the liquidity event. We therewith can determine the final wealth as follows:

$$\begin{aligned} W_1^i &= W_0^i + P_1 x_0^i \\ &= B_1^i + P_1 q_1^i, \end{aligned} \tag{4.3}$$

$$\begin{aligned} W_2^i &= B_1^i + P_2 q_1^i \\ &= B_2^i + P_2 q_2^i, \end{aligned} \tag{4.4}$$

$$W_3^i = B_2^i + P_3 q_2^i. \tag{4.5}$$

Inserting $B_2^i = B_1^i + P_2(q_1^i - q_2^i)$ and $B_1^i = W_0^i - P_1(q_1^i - x_0^i)$ from manipulating (4.3)

[7] See appendix A for an introduction into utility theory and the rationale behind the use of expected utility.

and (4.4), (4.5) becomes

$$
\begin{aligned}
(4.6) \qquad W_3^i &= B_1^i + P_2(q_1^i - q_2^i) + P_3 q_2^i \\
&= W_0^i + P_1(x_0^i - q_1^i) + P_2(q_1^i - q_2^i) + P_3 q_2^i \\
&= W_0^i + (P_2 - P_1)(q_1^i - x_0^i) + (P_3 - P_2)(q_2^i - x_0^i) + P_3 x_0^i.
\end{aligned}
$$

Define the excess demand over the liquidity event in period $t = 1, 2$ by investor i as

$$
(4.7) \qquad \xi_t^i = q_t^i - x_0^i.
$$

Replacing (4.7) in (4.6) yields

$$
(4.8) \qquad W_3^i = W_0^i + (P_2 - P_1)\xi_1^i + (P_3 - P_2)\xi_2^i + P_3 x_0^i.
$$

By using the approximation of the expected utility as presented in Appendix A.2 we obtain with risk aversion z^i:

$$
(4.9) \qquad E[U(W_3^i)] = U\left(E[W_3^i] - \frac{1}{2} z^i Var[W_3^i]\right).
$$

By using all available information at time $t = 2$, Ω_2, we get with (4.8)

$$
\begin{aligned}
(4.10) \qquad E[U(W_3^i)|\Omega_2] \;=\; U\Big(& W_0^i + (P_2 - P_1)\xi_1^i \\
& + (E[P_3|\Omega_2] - P_2)\xi_2^i + E[P_3|\Omega_2] x_0^i \\
& - \frac{1}{2} z^i (\xi_2^i + x_0^i)^2 Var[P_3|\Omega_2] \Big).
\end{aligned}
$$

Maximizing (4.10) for the optimal excess demand in period 2, ξ_2^i, gives the following first order condition:

$$
(4.11) \qquad \left(E[P_3|\Omega_2] - P_2 - z^i(\xi_2^i + x_0^i) Var[P_3|\Omega_2]\right) U'(\cdot) = 0.
$$

With $U'(\cdot) > 0$ and $U''(\cdot) < 0$ as required for risk averters the second order condition can easily be shown to be fulfilled. Solving for ξ_2^i gives the optimal excess demand in period 2 as

$$
(4.12) \qquad \xi_2^i = \frac{E[P_3|\Omega_2] - P_2}{z^i Var[P_3|\Omega_2]} - x_0^i.
$$

As only for investors we find that $x_0^i \neq 0$, for all other market participants the excess demand simplifies to

$$\xi_2^i = \frac{E[P_3|\Omega_2] - P_2}{z^i Var[P_3|\Omega_2]}, \tag{4.13}$$

where only market makers trade, hence this demand is only valid for market makers. For market participants that are neither investors nor market makers we find $\xi_2^i = 0$. Assuming all market participants to have the same risk aversion, i.e. $z^i = z$ for all $i = 1, \ldots, N$,[8] we can aggregate (4.12) and (4.13) and receive with (4.1):[9]

$$\xi_2^I = \sum_{i=1}^{n} \xi_2^i = n\frac{E[P_3|\Omega_2] - P_2}{zVar[P_3|\Omega_2]}, \tag{4.14}$$

$$\xi_2^M = \sum_{i=1}^{M} \xi_2^i = M\frac{E[P_3|\Omega_2] - P_2}{zVar[P_3|\Omega_2]}. \tag{4.15}$$

Market clearing in period 2 requires

$$0 = \xi_2^I + \xi_2^M = (n+M)\frac{E[P_3|\Omega_2] - P_2}{zVar[P_3|\Omega_2]}, \tag{4.16}$$

which implies

$$E[P_3|\Omega_2] = P_2 \tag{4.17}$$

unless $M = n = 0$.

By inserting (4.17) into (4.12) and (4.13) we obtain

$$\xi_2^i = -x_0^i \qquad \text{for investors,} \tag{4.18}$$

$$\xi_2^i = 0 \qquad \text{for market makers.} \tag{4.19}$$

With using (4.17) - (4.19) it is now possible to determine the optimal excess demand in period 1 by returning to (4.9) and evaluating this expression using all available

[8] Assuming equal risk aversion does not change the arguments to be derived below. We could easily proceed with different degrees of risk aversion at the cost of additional notation.

[9] It is necessary to aggregate over all investors in period 2 and not only over those trading in period 2, because the order imbalance of the investors having traded in period 1 has to be offset in period 2 by the market maker such that also the market maker holds his optimal portfolio at the end of the period.

information from period 1, Ω_1:

$$
\begin{aligned}
(4.20) \qquad E[U(W_3^i)|\Omega_1] &= U\Big(W_0^i + (E[P_2|\Omega_1] - P_1)\xi_1^i \\
&\quad + (E[P_3|\Omega_1] - E[P_2|\Omega_1])\xi_2^i + E[P_2|\Omega_1]x_0^i \\
&\quad - \frac{1}{2}z^i(\xi_1^i + x_0^i)^2 Var[E[P_3|\Omega_2]|\Omega_1]\Big) \\
&= U\Big(W_0^i + (E[P_3|\Omega_1] - P_1)\xi_1^i + E[P_3|\Omega_1]x_0^i \\
&\quad - \frac{1}{2}z^i(\xi_1^i + x_0^i)^2 Var[E[P_3|\Omega_2]|\Omega_1]\Big).
\end{aligned}
$$

The first order condition for a maximum is

$$
(4.21) \qquad \big(E[P_3|\Omega_1] - P_1 - z^i(\xi_1^i + x_0^i)Var[E[P_3|\Omega_2]\Omega_1]\big)\, U'(\cdot) = 0.
$$

The second order condition again can easily be proved to be fulfilled. Solving for ξ_1^i gives the optimal excess demand in period 1:

$$
(4.22) \qquad \xi_1^i = \frac{E[P_3|\Omega_1] - P_1}{z^i Var[E[P_3|\Omega_2]|\Omega_1]} - x_0^i.
$$

For market makers this again reduces to

$$
(4.23) \qquad \xi_1^i = \frac{E[P_3|\Omega_1] - P_1}{z^i Var[E[P_3|\Omega_2]|\Omega_1]}.
$$

Aggregating over all investors and market makers we get with (4.2) and the assumption of equal risk aversion for all market participants:

$$
(4.24) \qquad \xi_1^I = n_1 \frac{E[P_3|\Omega_1] - P_1}{zVar[E[P_3|\Omega_2]|\Omega_1]} - X_0^1,
$$

$$
(4.25) \qquad \xi_1^M = M \frac{E[P_3|\Omega_1] - P_1}{zVar[E[P_3|\Omega_2]|\Omega_1]}.
$$

Market clearing in period 1 requires

$$
(4.26) \qquad 0 = \xi_1^I + \xi_1^M = (n_1 + M)\frac{E[P_3|\Omega_1] - P_1}{z^i Var[E[P_3|\Omega_2]|\Omega_1]} - X_0^1.
$$

The rate of return of the market makers from their activity is given by

$$
(4.27) \qquad r = \frac{P_2 - P_1}{P_1} = \frac{E[P_3|\Omega_2] - P_1}{P_1}.
$$

With (4.26) the expected value and variance of this return is given by

$$Var[r|\Omega_1] = \frac{1}{P_1^2} Var[E[P_3|\Omega_2]|\Omega_1], \tag{4.28}$$

$$\begin{aligned} E[r|\Omega_1] &= \frac{E[E[P_3|\Omega_2]|\Omega_1] - P_1}{P_1} = \frac{E[P_3|\Omega_1] - P_1}{P_1} \\ &= \frac{X_0^1 z}{P_1(n_1+M)} Var[E[P_3|\Omega_2]|\Omega_1] \\ &= \frac{P_1 X_0^1 z}{n_1+M} Var[r|\Omega_1]. \end{aligned} \tag{4.29}$$

To derive the equilibrium number of market makers, i.e. the supply of immediacy, suppose that there exists a fixed cost of C for becoming a market maker, e.g. costs of the back office. The expected utility from not participating in the market for all market participants not facing a liquidity event is $E[U(W_0^i)|\Omega_1]$. The profits from acting as market maker are $(P_2 - P_1)\xi_1^i$, hence the expected utility is $E[U(W_0^i - C + (P_2 - P_1)\xi_1^i)|\Omega_1]$.

In equilibrium the expected utility from acting as market maker and not participating in the market have to be equal:

$$E[U(W_0^i)|\Omega_1] = E[U(W_0^i - C + (P_2 - P_1)\xi_1^i)|\Omega_1]. \tag{4.30}$$

As the risk from holding the initial portfolio is equal in both cases, it can be neglected in the further analysis and only the additional risk that arises from acting as market maker has to be considered.

Inserting (4.23) and (4.26) we get from (4.30):

$$\begin{aligned} E[U(W_0^i)|\Omega_1] &= E\left[U\left(W_0^i - C + (P_2 - P_1)\frac{E[P_3|\Omega_1] - P_1}{zVar[E[P_3|\Omega_2]|\Omega_1]}\right)\middle|\,\Omega_1\right] \\ &= E\left[U\left(W_0^i - C + (P_2 - P_1)\frac{X_0^1}{n_1+M}\right)\middle|\,\Omega_1\right] \\ &= U\left(W_0^i - C + (E[P_2|\Omega_1] - P_1)\frac{X_0^1}{n_1+M} - \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1]\right). \end{aligned}$$

By comparing coefficients we get

$$
\begin{aligned}
(4.31)\quad C &= (E[P_2|\Omega_1] - P_1)\frac{X_0^1}{n_1+M} - \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1] \\
&= (E[E[P_3|\Omega_2]|\Omega_1] - P_1)\frac{X_0^1}{n_1+M} - \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1] \\
&= (E[P_3|\Omega_1] - P_1)\frac{X_0^1}{n_1+M} - \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1] \\
&= \frac{P_1 X_0^1}{n_1+M}E[r|\Omega_1] - \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1] \\
&= z\left(\frac{X_0^1}{n_1+M}\right)^2 P_1^2 Var[r|\Omega_1] - \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1] \\
&= z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1] - \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1] \\
&= \frac{1}{2}z\left(\frac{X_0^1}{n_1+M}\right)^2 Var[P_2|\Omega_1].
\end{aligned}
$$

Solving for M as the number of market makers gives the equilibrium supply of immediacy:

$$
(4.32)\qquad M = |X_0^1|\sqrt{\frac{zVar[P_2|\Omega_1]}{2C}} - n_1.
$$

The number of market makers is larger the lower the costs of market making, the higher the order imbalance, the higher the risk aversion and uncertainty about the price in period 2 and the lower the number of investors trading in the first period. For a high number of market makers it would be optimal that only a few investors were trading in the first period having a large order imbalance.[10]

When determining the optimal number of market makers, the restriction $M \leq N-n$ has further to be taken into account. In this framework immediacy is provided because of the possible profits that can be derived from a temporary order imbalance. How large this order imbalance is and how many investors decide to trade in period 1 also depends on the costs charged by market makers for trading in the first period, i.e. the demand for immediacy. The higher these costs, the lower demand, i.e. $|X_0^1|$, will be.

[10] Grossman and Miller (1988, pp. 626) derive a similar result by assuming that the order imbalance is not known to the market makers in advance. For solving this problem they have to assume an exponential utility function.

4.1.2 Determination of the optimal market form

The last section presented a model to determine the number of market makers endogenously. It was assumed that only the order imbalance in periods 1 and 2 are offset by market makers. This implies that the total order flow at first is matched and only those orders that could not be matched in this process are routed to market makers. Implicitly therewith it has been assumed that this matching process is costless and takes place in an instant. In reality, however, this matching process will be costly. Searching for an offsetting order is time intensive and it is not ensured that an offsetting order can be found in a given time period, even if such an order exists. These costs can give rise to the emergence of a specialized market participant, the match maker. He can be used by investors to match their orders. As many investors choose this match maker, he can more easily find an offsetting order. Unlike the market maker, he does not offset the order by trading on his own account, but only matches two orders of investors. The match maker, as well as the market maker, makes it more easy to execute an order, hence they both facilitate trading and provide liquidity, what has been identified in chapter 2.2 to be one reason for the emergence of markets.

In chapter 4.1.1 it has been assumed that a market participant not facing a liquidity event only can choose to become a market maker or not to participate in the market. In this section a model will be presented in which a single market participant can choose to become a match maker or market maker.[11] Yavas (1992) provides a model how this market participant chooses between becoming a market maker or a match maker and hence whether an auction or a dealer market emerges.

We only have two groups of market participants, besides the match maker or market maker. One group assigns a value of P_1 and the other of P_2 to the asset. All market participants are risk neutral, implying that maximizing expected utility is equivalent

[11] The lack of competition between market makers or match makers does not effect the model to be presented here. We can similarly assume that all market participants have to make the same decision, i.e. that we are considering a group of market participants acting competitively.

to maximizing expected profits. Every investor knows the value he assigns to the asset, but not the value the other group assigns. Their value is a random variable whose distribution function $F_i(P_i)$ is known to all market participants.

There is not a fixed buyer and a fixed seller group. If the price used for a transaction is below the own value assigned to the asset, it is bought, otherwise it is sold.[12] If two investors meet, they reveal their values of the asset. The joint surplus of a trade, $|P_1 - P_2|$, is assumed to be shared equally, each investor receiving $\frac{1}{2}|P_1 - P_2|$.[13] For the following we assume without loss of generality that $P_1 > P_2$.

Let further $0 \leq \theta \leq 1$ denote the probability at which two investors of different groups meet without a match maker or market maker. This probability depends on the search intensity of the two investors, A_1 and A_2, i.e. $\theta = \theta(A_1, A_2)$. We assume this probability to be increasing and concave in the search intensities of both groups. We further assume the costs to increase faster than the probability to meet and that investors not searching face no costs:

$$\begin{aligned} \frac{\partial \theta(A_1, A_2)}{\partial A_i} &> 0, \\ \frac{\partial^2 \theta(A_1, A_2)}{\partial A_i^2} &< 0 \qquad i = 1, 2, \\ \theta(A_1, 0) = \theta(0, A_2) &= 0. \end{aligned} \tag{4.33}$$

The costs of searching for investor i are denoted $C_i(A_i)$, and are increasing and

[12] If the two groups assign the same value to the asset no transaction occurs, but for simplicity we neglect this case without changing results.

[13] Yavas (1992, p. 36) shows that other divisions of the surplus do not affect the main results of the model.

convex in the search intensities of the corresponding group:

$$
\begin{aligned}
\frac{\partial C_i(A_i)}{\partial A_i} &> 0 \qquad i = 1,2, \\
\frac{\partial^2 C_i(A_i)}{\partial A_i^2} &> 0 \qquad i = 1,2, \\
\frac{\partial C_i(A_i)}{\partial A_{3-i}} &= 0 \qquad i = 1,2, \\
\frac{\partial C_i(A_i)}{\partial A_i} &> \frac{\partial \theta(A_1, A_2)}{\partial A_i} \qquad i = 1,2, \\
C_i(0) &= 0.
\end{aligned} \tag{4.34}
$$

The functions $\theta(A_1, A_2)$ and $C_i(A_i)$ are known by all market participants. As for an investor the value his trading partner assigns to the asset is random, we get the expected surplus from a trade in the absence of a market or match maker:

$$
S_i(P_i) = \int_0^\infty \frac{1}{2} |P_i - P_{3-i}| \, dF_{3-i}(P_{3-i}). \tag{4.35}
$$

The search intensity of the other group of investors is not known to investors, hence they have to be conjectured. This conjecture is denoted A^0_{3-i}. The expected profits from searching are

$$
\pi_i(A_i, P_i, A^0_{3-i}) = \theta(A_i, A^0_{3-i}) S_i(P_i) - C_i(A_i). \tag{4.36}
$$

Maximizing expected profits by choosing an optimal search intensity gives the following first order condition:

$$
\begin{aligned}
\frac{\partial \pi_i(A_i, P_i, A^0_{3-i})}{\partial A_i} &= \frac{\partial \theta(A_i, A^0_{3-i})}{\partial A_i} S_i(P_i) - \frac{\partial C_i(A_i)}{\partial A_i} \\
&= 0.
\end{aligned} \tag{4.37}
$$

The second order condition for a maximum is fulfilled if $\frac{\partial^2 C_i(A_i)}{\partial A_i^2} > \frac{\partial^2 \theta(A_i, A^0_{3-i})}{\partial A_i^2} S_i(P_i)$. The term on the left side is positive from assumption (4.34), the first term on the right side is negative from (4.33) and the second term positive as can be seen from its definition in (4.35), hence the left side is negative. The second order condition therewith is always fulfilled with the assumptions stated above.

As we assume all functions to be common knowledge, investors can correctly infer the search intensity of the investors of the other group by solving their equation

(4.37). Inserting this result, A^*_{3-i}, they can solve (4.37) to find their own optimal search activity.

Thus far we have not introduced a match maker or market maker, investors are supposed to search for each other on their own behalf. Without introducing such an intermediaire this search equilibrium is subject to two inefficiencies. The first inefficiency is that investors may not meet, although they would like to trade with each other. The second inefficiency is that the probability of investors meeting depends on the search activities of both groups, hence each group produces positive externalities to the other group. Therefore in equilibrium the search activity will be lower than in the social optimum.

These inefficiencies enable match or market makers to be established in the market. It will first be analyzed how investors behave in such markets and then these two market forms are compared.

We first analyze the case where the intermediaire is a market maker. Like the investors we assume him to be risk neutral. He is also assumed to know the distributions of the fundamental values each group of investors assigns to the asset, but not the valuation itself. Furthermore he knows the cost functions for searching and the probability function of meeting upon searching.

At the beginning of the period a market maker quotes his prices at which he is willing to sell the asset, P_a, and at which he is willing to buy the asset, P_b, where $P_a \geq P_b$.[14] These prices become known to all investors at no costs. They then have to decide whether they want to search for an offsetting order on their own or whether they want to trade directly with the market maker at the stated prices. By searching on their own, investors hope to receive a better price than the quotes of the market maker. If the search is not successful, i.e. they find no offsetting order, or upon meeting the price is less favorable than the quotes of the market maker,

[14] If $P_a < P_b$ an investor could buy the asset at P_a from the market maker and sell it to him in an instant of time at P_b making a profit, while the market maker would make a loss. To prevent this arbitrage it is required that $P_a \geq P_b$. Chapters 4.3 and 4.4 present more sophisticated models of market making showing that in general $P_a > P_b$.

they can yet trade with the market maker at the same stated prices. But, as the search process takes time, this trade will take place at the end of the period, hence the surplus achieved has to be discounted by a rate of ρ.[15]

If $P_i > P_a$ the investor will buy the asset from the market maker and if $P_i < P_b$ he will sell it to him. If $P_b > P_i > P_a$ the investor would make a loss from trading with the market maker and hence will not trade with him.

The surplus from trading with the market maker, $P_i - P_a$ and $P_b - P_i$, respectively, is now not divided as the market maker quotes firm prices. In the presence of a market maker the surplus from a direct trade between investors has to be divided as follows: to prevent the investors to refuse the trade and trade with the market maker instead, they have at first to receive the discounted surplus they would get from trading with him afterwards. After both investors have received this compensation for the surplus they could receive from trading with the market maker afterwards, the remaining surplus is divided equally between them. If the surplus from a direct trade with each other cannot give this minimum compensation, it is refused. They only trade with each other if

$$P_1 - P_2 \geq \rho \left(\max\{0, P_1 - P_a, P_b - P_1\} + \max\{0, P_2 - P_a, P_b - P_2\} \right). \tag{4.38}$$

Without an intermediaire the reservation prices of the investors have been their valuation of the asset. With the existence of a market maker these reservation prices change. For trading with each other, the price of an investor to buy the asset from another has to be reduced by the surplus he could earn from trading with the market maker. Similarly for the investor selling the asset the reservation price is increased by this amount. The reservation prices become

$$\begin{aligned} P_1^r &= P_1 - \rho \max\{0, P_1 - P_a, P_b - P_1\}, \\ P_2^r &= P_2 + \rho \max\{0, P_2 - P_a, P_b - P_2\}. \end{aligned} \tag{4.39}$$

[15] This discount factor could also be interpreted as an adapting utility to the risk that prices have changed unfavorably in the mean time if we assume risk averse market participants.

The negotiated price will always be between these two reservation prices. If $P_1^r \geq P_2^r$ we see that

$$P_1 \geq P_1^r \geq P_2^r \geq P_2. \tag{4.40}$$

The prices are allowed to change only in a smaller interval by subsequent trades in presence of a market maker, hence the price dispersion is reduced.

The additional surplus from trading directly with each other instead with the market maker is $\frac{1}{2}(P_1^r - P_2^r)$ if $P_1^r > P_2^r$, otherwise no trade occurs. There also does not occur a trade if the investors do not meet. They will not meet if one of them decides to trade with the market maker directly instead of searching first, i.e. his search activity is zero ($A_i = 0$). Therefore we define

$$I_i^D(A_i) = \begin{cases} 1 & \text{if } A_i > 0 \\ 0 & \text{if } A_i = 0 \end{cases}. \tag{4.41}$$

The expected surplus from a trade with each other now becomes

$$S_i^D(P_i, P_a, P_b) = \int_0^\infty \frac{1}{2} \max\{0, P_1^r - P_2^r\} I_{3-i}^D(A_{3-i}) dF_{3-i}(P_{3-i}) \tag{4.42}$$

As we see from (4.40) and (4.41) that $P_1^r - P_2^r \leq P_1 - P_2$ and $I_{3-i}^D(A_{3-i})F_i(P_i) \leq F_i(P_i)$, it is obvious that the expected surplus from a trade with each other is reduced in the presence of a market maker:

$$S_i(P_i) \geq S_i^D(P_i, P_a, P_b). \tag{4.43}$$

The expected profits are determined by the expected surplus from trading with each other, the costs of searching,[16] and the expected surplus from trading with the market maker:

$$\begin{aligned} \pi_i(A_i, P_i, A_{3-i}^0, P_a, P_b) &= S_i^D(P_i, P_a, P_b) - C_i(A_i) \\ &\quad + (1 - \theta(A_i, A_{3-i}^0))\rho \max\{0, P_i - P_a, P_p - P_i\}. \end{aligned} \tag{4.44}$$

[16] These costs also have to be beared if no trade occurs because they are sunk costs.

Maximizing to determine the optimal search activity, A_i, gives the following first order condition:[17]

$$
\begin{aligned}
(4.45) \qquad 0 &= \frac{\partial \pi_i(A_i, P_i, A^0_{3-i}, P_a, P_b)}{\partial A_i} \\
&= \frac{\partial \theta(A_i, A^0_{3-i})}{\partial A_i} \left(S^D_i(P_i, P_a, P_b) - \rho \max\{0, P_i - P_a, P_b - P_i\} \right) \\
&\quad - \frac{\partial C_i(A_i)}{\partial A_i}.
\end{aligned}
$$

By inspection of (4.45) we see that the term in brackets has to be smaller than $S_i(P_i)$, but still is positive. Let us for notational simplicity rewrite (4.37) and (4.45) as

$$\theta' S_i - C'_i = 0, \tag{4.46}$$

$$\theta'_D S^{D'}_i - C'_D = 0, \tag{4.47}$$

where $S^{D'}_i = S^D_i(P_i, P_a, P_b) - \rho \max\{0, P_i - P_a, P_b - P_i\} \leq S_i$. Solving for S_i and $S^{D'}_i$ we get

$$
\begin{aligned}
S_i &= \frac{C'_i}{\theta'}, \\
S^{D'}_i &= \frac{C'_D}{\theta'_D}.
\end{aligned}
$$

hence we have

$$\frac{C'_i}{\theta'} \geq \frac{C'_D}{\theta'_D}. \tag{4.48}$$

From (4.34) we can find a k and a k_D such that $C'_i = k\theta'$ and $C'_D = k_D \theta'_D$, where because of (4.33) and (4.34) k and k_D are strictly increasing in A_i. Inserting these relations into (4.48) we obtain

$$k \geq k_D. \tag{4.49}$$

Hence the optimal search activity is smaller in the presence of a market maker than without. This is the result of the possibility to trade with the market maker instead of only trading directly with other investors.

[17] The second order condition can be proved to be fulfilled in the same manner as in (4.37).

The optimal search intensity can be determined using (4.45). Based on these inferences of the optimal search activities the market maker can set his prices optimal to maximize his expected profits. YAVAS (1992, pp. 42 ff.) shows that an overall equilibrium exists and addresses the problems of multiple equilibria that are of no importance for our purpose. It can easily be proved that the expected profits of the investors are higher in the presence of a market maker, hence it is beneficial for investors to have a market maker. A market maker will also make profits from his activity and therefore market participants will act as market makers.

We now turn to the case where the intermediaire has decided to become a match maker. Like in the presence of a market maker, investors can either decide to search for an offsetting order by themselves, facing the risk that they do not meet, or use the match maker where we assume that he has the certainty that the order will be executed.[18] If they decide to search and do not find an offsetting order, they can afterwards turn to the match maker for execution and receive the discounted surplus. For his service the match maker charges a fee from both investors whose orders he matches, the fee is a fraction c of the transaction price. The price is determined such by the match maker that the surplus is equally distributed between investors.

The reservation prices for a direct trade with each other are again the valuations of the asset as no other market participant offers a better price. The match maker only matches the orders and offers no fixed price. The surplus from trading will be $P_1 - P_2$, which will be divided equally among investors. When trading with the help of a match maker the surplus first has to cover the fees that have to be paid to the match maker. If the surplus does not cover these costs, no trade will occur. With P denoting the transaction price, we get

$$\begin{aligned} P_1 - P &\geq cP, \\ P - P_2 &\geq cP, \end{aligned} \tag{4.50}$$

[18] We could also assume that the probability of execution is significantly higher by using the match maker without changing the argument.

which implies

$$
\begin{aligned}
P &\leq P_1^r = \frac{P_1}{1+c} \leq P_1, \\
P &\geq P_2^r = \frac{P_2}{1-c} \geq P_2,
\end{aligned} \tag{4.51}
$$

where P_i^r denotes the reservation price for trading with the help of a match maker. A trade through the match maker will only occur at prices between P_2^r and P_1^r, with $P_1^r > P_2^r$. As this interval is smaller than $[P_2, P_1]$ the price dispersion is reduced like in the presence of a market maker. The surplus from trading with a match maker is given by

$$
\begin{aligned}
P_1 - P_2 - c(P_1 + P_2) &= (1+c)P_1^r - (1-c)P_2^r \\
&\quad -c\left((1+c)P_1^r + (1-c)P_2^r\right) \\
&= P_1^r(1+c)(1-c) - P_2^r(1+c)(1-c) \\
&= (1-c^2)(P_1^r - P_2^r) \\
&> 0.
\end{aligned} \tag{4.52}
$$

If $P_2^r > P_1^r$ the surplus would be negative and no trade would occur. Like in the presence of a market maker, investors can also directly use the match maker instead of searching for an offsetting order by themselves, i.e. $A_i = 0$. Hence we define

$$
I_i^M(A_i) = \begin{cases} 1 & \text{if } A_i > 0 \\ 0 & \text{if } A_i = 0 \end{cases}. \tag{4.53}
$$

The expected surpluses from trading directly with each other and by using the match maker are given by

$$
S_i(P_i) = \int_0^\infty \frac{1}{2} |P_i - P_{3-i}| I_{3-i}^M(A_{3-i}) dF_{3-i}(P_{3-i}), \tag{4.54}
$$

$$
S_i^M(P_i, c) = \int_0^\infty \frac{1}{2} \max\{0, |P_i - P_{3-i}| - c(P_1 + P_2)\} I_{3-i}^M(A_{3-i}) dF_{3-i}(P_{3-i}). \tag{4.55}
$$

therewith the expected profit of an investor trading with the match maker is

$$
\begin{aligned}
\pi_i^M(P_i, A_i, A_{3-i}^M, c) &= \theta(A_i, A_{3-i}^0) S_i(P_i) \\
&\quad + (1 - \theta(A_i, A_{3-i}^0)) \rho S_i^M(p_i, c) - C_i(A_i).
\end{aligned} \tag{4.56}
$$

Maximizing the expected profits to find the optimal search activity gives the following first order condition:

$$
\begin{aligned}
(4.57) \qquad 0 &= \frac{\partial \pi_i(P_i, A_i, A_{3-i}^M, c)}{\partial A_i} \\
&= \frac{\partial \theta(A_i, A_{3-i}^0)}{\partial A_i}(S_i(P_i) - \rho S_i^M(P_i, c)) - \frac{\partial C_i(A_i)}{\partial A_i}.
\end{aligned}
$$

As from inspection of (4.54) and (4.55) we see that $S_i(P_i) \geq S_i^M(P_i, c) \geq \rho S_i^M(P_i, c) \geq 0$, it can easily be verified that $0 \leq S_i(P_i) - \rho S_i^M(P_i, c) \leq S_i(P_i)$. This is in line with the results in the presence of a market maker, the second order condition is fulfilled and it turns out that the search activities are reduced in the presence of a match maker. The optimal search activity can be derived from (4.57) and given these results the match maker can determine his optimal fee. YAVAS (1992, pp. 50 f.) again shows the existence of such an equilibrium. Like with the presence of a market maker, the expected profits of the investors are higher in presence of a match maker, hence it is beneficial for investors to have a match maker. A match maker will also make profits from his activity and therefore there will be market participants acting as match maker.

Thus far it has only been shown that a market maker as well as a match maker are beneficial for investors and that both forms will be provided as they make a profit from their activities. Nothing has been said about which form will be preferred. We therefore now turn to the choice a market participant will make, whether to become a market maker or a match maker. The decision is made on the basis of the expected returns one can earn from these two activities. YAVAS (1992, pp. 51 ff.) provides some conditions for this choice.

A market participant will decide to become a market maker if the value he assigns to the asset differs largely from that of the investors, i.e. if his valuation is at the tail of the distribution of the investors. By offering a very favorable price to both investor groups he still makes a sufficient profit from his activity with a high probability as it is very unlikely that both investors will find it more profitable to trade with each other even if they meet. If search is very efficient, i.e. the costs are relatively small,

two investors searching are very likely to meet, but they in most cases will not trade with each other as the market maker can offer a more favorable price and even with very efficient search the market maker will make a considerable profit.

However, if the valuation of the asset does not differ much between investors and the intermediaire and search is inefficient, i.e. imposes high costs, he will choose to become a match maker. As market maker he would in many cases not offer a more favorable price to the investors than if they trade directly with each other, so that he is unlikely to serve an order, hence his expected profits are low. Whereas a match maker charges fees from both sides independent of the price he charges, giving him higher profits. If search is efficient, the decision whether to become a match maker or market maker will depend on the exact parameter constellation.

These conditions to become a market maker or a match maker may change over time and therefore he may want to revise his decision. As a market should apply the same rules over a longer period of time to give investors stable environments to trade and enable the price mechanism to work properly, frequent changes of the market forms should not be allowed. When taking into account that a decision binds him over a given period the above criteria have to be made in a dynamic setting, taking into account changes that are expected. The general idea is not affected, it has to be taken the (discounted) profits over the time period he is bound.

The decision in reality often is restricted by rules stating that all assets in a certain market have to be traded with the same rules. For the determination of the optimal market form in this case the costs and benefits have to aggregated. A certain variety can be achieved by defining market segments that can be traded according to different rules. The assets are then traded into one of these segments according to their characteristics.

4.2 Auction markets

This section will give an overview of the main contributions in market microstructure theory on auction markets. Prices are not analyzed on a trade-by-trade basis, but aggregated over a given period of time, e.g. several hours of a trading day or an entire trading day. These models allow to explain how information is incorporated into prices.

We assume that a single asset is traded in T trading rounds by two groups of investors, informed and uninformed investors. After the final round of trading the asset is liquidated and the proceedings are consumed. Risk neutral informed investors, called *insiders*,[19] receive a perfect signal on the liquidation value of the asset before the first round of trading. The uninformed investors do not receive a signal, hence their trades are not based on superior information, but on exogenous needs for liquidity (*noise traders*) as in the first two models or they rebalance their portfolios as a result of portfolio imbalances (*hedgers*) imposed exogenously in the final model.

In each trading round (auction) investors can submit their orders to a match maker.[20] The match maker arranges the trades at a single price. This price has to be set such that it equals the expected liquidation value given the match makers' information, hence he makes zero expected profits and therewith is assumed to behave competitively. The information the match maker has, as he is assumed to be uninformed, is the order imbalance of that and former auctions.

[19] The term *insider* as used in this context has to be distinguished from the legal definition of an insider. Here an insider is an investor having acquired information, that is not known to all market participants. In the legal definition an insider has access to not yet public available information due to his position in a company, e.g. as member of the board. See BÜSCHGEN (1991, p.362). Legislation in all advanced countries prohibits the use of such information for trading.

[20] We allow only for market orders, the submission of limit orders, i.e. demand schedules, is prohibited. When submitting an order, the orders submitted by other investors for this auction are not known to any investor, i.e. the choices have to be simultaneous and cannot be conditioned on the behavior of other investors.

As the match maker does not know whether the orders submitted come from informed or uninformed investors (the use of brokers ensures anonymity of investors to the match maker), more orders to buy than to sell could mean that the informed investors received a signal indicating the liquidation value to be above the last observed price. But it could also be the result of pure incident due to noise trader behavior. This price setting behavior is identical to the efficient market hypothesis, where the price equals the expected value of the asset given a set of information.[21] For simplicity it is assumed that the match maker charges no fee for his service.

The next section presents the KYLE (1985) model. Although this model is included as a special case in section 4.2.2, it is treated separately due to its outstanding position in the development of theories on auction markets. It assumes a single informed investor and a large number of noise traders. In section 4.2.2 this model is extended to include more than one informed investor. A last extension in 4.2.3 assumes that the uninformed investors are no longer noise traders, but risk averse hedgers maximizing their own utility. In section 4.2.4 we determine the information revealed by trading volume. Finally in 4.2.5 it is shown how the developed models can be used to explain some effects observed in stock markets. The presentation of the models has throughout been adapted to use the same framework for all models and concentrate on the most important aspects.

4.2.1 Auctions with a single informed investor

KYLE (1985) presents a model where a single informed investor trades a single asset together with N uninformed noise traders. It is assumed that the informed investor becomes to know the liquidation value v of the asset with certainty.[22] For

[21] Although this behavior is very restrictive, it captures some of the behaviors in real stock exchanges applying auction markets. For example the Frankfurt Stock Exchange urges the match makers to rise the price if more buy than sell orders are in the market. See DEUTSCHE BÖRSE GROUP (1999, Part 6, 3.22 and 3.3.1.2.).

[22] The original model assumes that the signal is not perfect, but that the liquidation value has a positive variance given this information. As informed investors are assumed to be risk neutral this remaining risk has not to be considered in the optimal behavior as maximizing expected profits and expected utility are equivalent.

uninformed investors the liquidation value is a random variable v that is normally distributed with initial mean p_0 and variance Σ_0:

$$v \sim N(p_0, \Sigma_0). \tag{4.58}$$

Uninformed investors are assumed to trade for purely exogenous reasons, they do not maximize any objective function. Their order sizes, u^i, are random variables that are supposed to be independently identically normally distributed with mean zero and variance $\sigma^2_{u^i}$.[23] They are independent of the order sizes of other uninformed investors, the behavior of the informed investor, independent over time and of v:[24]

$$u^i \sim N\left(0, \sigma^2_{u^i}\right). \tag{4.59}$$

The total relevant order flow from uninformed investors is their order imbalance, all other orders can directly be matched:

$$u = \sum_{i=1}^{N} u^i \sim N\left(0, N\sigma^2_{u^i}\right) = N\left(0, \sigma^2_u\right). \tag{4.60}$$

We assume that T, N, $\sigma^2_{u^i}$, p_0 and Σ_0 are known by all market participants as well as that the variables are independently normally distributed.

The order size of the informed investor is denoted x. The match maker cannot distinguish between orders from informed and uninformed investors, hence he only can observe the aggregated order flow $u + x$. This order flow is used to determine the price according to

$$p = E[v|u + x]. \tag{4.61}$$

Kyle (1985) first investigates the optimal behavior of the informed investor by choosing an optimal x in a single auction, i.e. for $T = 1$. The informed investor

[23] In deriving the results the assumption of normality is central. Relaxing this assumption to the class of elliptical functions gives similar results, but further generalizations may give different results. See O'Hara (1995, p. 99).

[24] Obviously, noise traders submit orders independent of trading costs, i.e. the charged price. Ramanlal (1999) shows that such a behavior does not violate the assumption of rational investors under conditions that are fulfilled in real markets.

maximizes his expected profits from trading, given his information on the liquidation value of the asset. His profits are

$$\pi = (v - p)x. \tag{4.62}$$

Only linear equilibria are considered by KYLE (1985). This assumption gives rise to the following pricing and order submission rules:

$$p = \mu + \lambda(x + u), \tag{4.63}$$

$$x = \alpha + \beta v. \tag{4.64}$$

With (4.62) - (4.64) we find the expected profits of the informed investor to be

$$\begin{aligned} E[\pi|v] &= E[(v - \mu - \lambda(x + u))x|v] \\ &= (v - \mu - \lambda x)x. \end{aligned} \tag{4.65}$$

Maximizing (4.65) to determine the optimal order size of the informed investor gives the following first order condition:

$$v - \mu - \lambda x - \lambda x = v - \mu - 2\lambda x = 0. \tag{4.66}$$

Rearranging yields

$$x = -\frac{\mu}{2\lambda} + \frac{v}{2\lambda}. \tag{4.67}$$

Comparing coefficients with (4.64) we get

$$\begin{aligned} \beta &= \frac{1}{2\lambda}, \\ \alpha &= -\frac{\mu}{2\lambda} = -\mu\beta. \end{aligned} \tag{4.68}$$

The second order condition for a maximum,

$$-2\lambda < 0, \tag{4.69}$$

states that we only have to consider positive λ. Using (4.61) we get with (4.63) and (4.64) and the results of the conditional mean of jointly normal distributed random

variables:

$$
\begin{aligned}
(4.70) \qquad p &= E[v|x+u] \\
&= E[v] + \frac{Cov[v, x+u]}{Var[x+u]}(x+u-E[x+u]) \\
&= p_0 + \frac{\beta\Sigma_0}{\beta^2\Sigma_0+\sigma_u^2}(\alpha+\beta p_0) + \frac{\beta\Sigma_0}{\beta^2\Sigma_0+\sigma_u^2}(x+u).
\end{aligned}
$$

By comparing coefficients with (4.63) we see that

$$
\begin{aligned}
(4.71) \qquad \lambda &= \frac{\beta\Sigma_0}{\beta^2\Sigma_0+\sigma_u^2}, \\
\mu &= p_0 - \frac{\beta\Sigma_0}{\beta^2\Sigma_0+\sigma_u^2}(\alpha+\beta p_0) = p_0 - \lambda(\alpha+\beta p_0).
\end{aligned}
$$

Solving (4.68) and (4.71) we get with (4.69):

$$
\begin{aligned}
(4.72) \qquad \beta &= \sqrt{\frac{\sigma_u^2}{\Sigma_0}}, \\
\lambda &= 2\sqrt{\frac{\Sigma_0}{\sigma_u^2}}, \\
\mu &= p_0, \\
\alpha &= -\beta p_0.
\end{aligned}
$$

As a result we can rewrite (4.63) and (4.64) as

$$
(4.73) \qquad p = p_0 + \lambda(x+u),
$$

$$
(4.74) \qquad x = \beta(v-p_0).
$$

The linear equilibrium exists and is unique. Nothing can be said about the existence of further nonlinear equilibria.

Uninformed investors cannot observe the order flow, only the price that is set by the match maker. Using this information they can update their beliefs on the distribution of the liquidation value. It is easy to show that[25]

$$
\begin{aligned}
(4.75) \qquad \Sigma_1 &= Var[v|p] \\
&= Var[v] - \frac{Cov[v,p]^2}{Var[p]} \\
&= \frac{1}{2}\Sigma_0,
\end{aligned}
$$

[25] See GREENE (1997, p. 90) for the determination of the conditional expected value and conditional variance of jointly normal distributed random variables.

$$
\begin{aligned}
p_1 &= E[v|p] \\
&= E[v] + \frac{Cov[v,p]}{Var[p]}(p - E[p]) \\
&= p + 2\lambda^2 u.
\end{aligned} \tag{4.76}
$$

As $E[u|p] = 0$ we find $E[p_1|p] = p$, hence the posterior distribution of v is

$$
v|p \sim N\left(p, \frac{1}{2}\Sigma_0\right). \tag{4.77}
$$

The variance uninformed investors attribute to the liquidation value of the asset can be interpreted as how much information is incorporated into the price. A variance of zero has to be interpreted as a perfect revelation of the information through prices, the closer this variance is to Σ_0 the less informative the price is. As by observing the price set by the match maker, the variance halfs, it can be said that half of the information is incorporated into prices. The variance of the liquidation value we can view as a measure for the *informativeness of prices*. The smaller the conditional variance the more informative prices are.

λ measures the influence an additional unit of an order has on the price as can be derived from (4.73):

$$
\frac{\partial p}{\partial(x+u)} = \lambda. \tag{4.78}
$$

We define a market as liquid if by placing an additional order the price does not change. Therefore λ measures the liquidity of a market, the closer it is to zero the more liquid the market is.[26] Usually $1/\lambda$ is taken as a measure of liquidity as by this definition a larger value corresponds to higher liquidity.

From (4.72) we see that a market is more liquid the more liquidity traders are in the market, i.e. the larger N is, or the more their order sizes vary, i.e the larger $\sigma^2_{u^i}$ is.

This framework can now be extended to trades occurring in $1 < T < \infty$ auctions in a given time period $[0, 1]$. This setting allows the informed investor to time his

[26] See KYLE (1985, pp. 1316 ff.).

trades such that over time his expected profits are maximized. The order flow from an uninformed investor for auction $1 \leq t \leq T$ is denoted u_t^i. The variance of the order flow from uninformed investors, $\sigma^2_{u^i}$, is assumed to remain constant over the entire period, i.e.

$$\sigma^2_{u^i} = \sum_{t=1}^{T} \sigma^2_{u^i,t} = T\sigma^2_{u^i,t}. \tag{4.79}$$

Hence we have

$$u_t^i \sim N\left(0, \frac{1}{T}\sigma^2_{u^i}\right) \tag{4.80}$$

for all $t = 1, \ldots, T$. The total order flow from the informed investor in the first t auctions is denoted x_t and the order flow for a specific auction k by Δx_k. As a result we have for all $t = 1, \ldots, T$:

$$x_t = \sum_{k=1}^{t} \Delta x_k. \tag{4.81}$$

When again considering only linear equilibria we get in analogy to (4.63) and (4.64) for all $t = 1, \ldots, T$:

$$p_t = \mu_t + \lambda_t(\Delta x_t + u_t), \tag{4.82}$$

$$\Delta x_t = \alpha_t + \beta_t v. \tag{4.83}$$

The match maker sets his price again such that it equals the liquidation value of the asset given the order flows observed in the past:

$$p_t = E[v|\Omega_t], \tag{4.84}$$

where $\Omega_t = \{\Delta x_1 + u_1, \ldots, \Delta x_t + u_t\}$. The informed investor maximizes his expected profits by choosing the optimal order size given the liquidation value and past prices. His profits from the remaining auctions are given by[27]

$$\pi_t = \sum_{k=t}^{T} (v - p_k)x_k = (v - p_t)\Delta x_t + \pi_{t+1}. \tag{4.85}$$

[27] It is worth noting that it is assumed here that future profits are not discounted to their present value.

We assume the expected profits to be quadratic in $v - p_t$:

$$E[\pi_{t+1}|p_1, \ldots, p_t, v] = \gamma_t(v - p_t)^2 + \delta_t. \tag{4.86}$$

Using (4.85) we get as the objective function of the informed investor:

$$\begin{aligned} E[\pi_{t+1}|p_1, \ldots, p_t, v] &= E[(v - p_t)\Delta x_t + \pi_{t+1}|p_1, \ldots, p_t, v] \\ &= E[(v - p_t)\Delta x_t|p_1, \ldots, p_t, v] \\ &\quad +\gamma_t(v - p_t)^2 + \delta_t \\ &= (v - \mu_t - \lambda_t \Delta x_t)\Delta x_t + \delta_t + \gamma_t \lambda_t^2 \sigma_{u^i,t}^2 \\ &\quad +\gamma_t(v - \mu_t - \lambda_t(\Delta x_t + u_t))^2. \end{aligned} \tag{4.87}$$

Maximizing this expression for the optimal order size to submit in auction t, Δx_t, gives the following first order condition:

$$\begin{aligned} 0 &= v - \mu_t - \lambda_t \Delta x_t - \lambda_t \Delta x_t + 2\gamma_t(v - \mu_t - \lambda_t \Delta x_t)(-\lambda_t) \\ &= (v - \mu_t)(1 - 2\gamma_t\lambda_t) - 2\lambda_t \Delta x_t(1 - \gamma_t\lambda_t). \end{aligned} \tag{4.88}$$

Solving for Δx_t we get

$$\Delta x_t = -\frac{1 - 2\gamma_t\lambda_t}{2\lambda_t(1 - \gamma_t\lambda_t)}\mu_t + \frac{1 - 2\gamma_t\lambda_t}{2\lambda_t(1 - \gamma_t\lambda_t)}v. \tag{4.89}$$

Comparing coefficients with (4.83) gives

$$\beta_t = \frac{1 - 2\gamma_t\lambda_t}{2\lambda_t(1 - \gamma_t\lambda_t)}, \tag{4.90}$$

$$\alpha_t = -\frac{1 - 2\gamma_t\lambda_t}{2\lambda_t(1 - \gamma_t\lambda_t)}\mu_t = -\beta_t\mu_t. \tag{4.91}$$

The second order condition

$$-2\lambda_t(1 - \gamma_t\lambda_t) < 0 \tag{4.92}$$

implies that an equilibrium exists only if $\lambda_t(1 - \gamma_t\lambda_t) > 0$. Using (4.84) we get

$$\begin{aligned} p_t &= E[v|\Omega_t] \\ &= E[E[v|\Omega_{t-1}]|\Delta x_t + u_t] \\ &= p_{t-1} + \frac{\beta_t \Sigma_{t-1}}{\beta_t^2 \Sigma_{t-1} + \sigma_{u^i,t}^2}(\Delta x_t + u_t - (\alpha_t + \beta_t p_{t-1})). \end{aligned} \tag{4.93}$$

Comparing coefficients with (4.82) we receive

$$\begin{aligned}(4.94)\qquad \lambda_t &= \frac{\beta_t \Sigma_{t-1}}{\beta_t^2 \Sigma_{t-1} + \sigma^2_{u^i,t}},\\ \mu_t &= p_{t-1} - \frac{\beta_t \Sigma_{t-1}}{\beta_t^2 \Sigma_{t-1} + \sigma^2_{u^i,t}}(\alpha_t + \beta_t p_{t-1})\\ &= p_{t-1} - \lambda_t(\alpha_t + \beta_t p_{t-1}).\end{aligned}$$

Solving (4.90) and (4.94) this becomes

$$\begin{aligned}(4.95)\qquad \lambda_t &= \frac{\beta_t \Sigma_{t-1}}{\beta_t^2 \Sigma_{t-1} + \sigma^2_{u^i,t}},\\ \beta_t &= \frac{1 - 2\lambda_t\gamma_t}{2\lambda_t(1 - \lambda_t\gamma_t)},\\ \mu_t &= p_{t-1},\\ \alpha_t &= -\beta_t p_{t-1}.\end{aligned}$$

This leads to a unique linear equilibrium:

$$(4.96)\qquad p_t = p_{t-1} + \lambda_t(\Delta x_t + u_t),$$

$$(4.97)\qquad \Delta x_t = \beta_t(v - p_{t-1}).$$

By comparing the coefficients in (4.87) and (4.86) we see that

$$(4.98)\qquad \delta_{t-1} = \delta_t + \gamma_t \lambda_t^2 \sigma^2_{u^i,t},$$

$$\begin{aligned}(4.99)\qquad \gamma_{t-1}(v - p_{t-1}) &= (v - \mu_t - \lambda_t \Delta x_t)\Delta x_t + \gamma_t(v - \mu_t - \lambda_t \Delta x_t)^2\\ &= \left(v - p_{t-1} - \lambda_t\beta_t(v - p_{t-1})\right)\beta_t(v - p_{t-1})\\ &\quad + \gamma_t\left(v - p_{t-1} - \lambda_t\beta_t(v - p_{t-1})\right)^2\\ &= (v - p_{t-1})^2(1 - \lambda_t\beta_t)\beta_t\\ &\quad + \gamma_t(v - p_{t-1})^2(1 - \lambda_t\beta_t)^2\\ &= (v - p_{t-1})^2\left((1 - \lambda_t\beta_t)\beta_t + \gamma_t(1 - \lambda_t\beta_t)^2\right).\end{aligned}$$

hence

$$\begin{aligned}(4.100)\qquad \gamma_{t-1} &= (1 - \lambda_t\beta_t)\beta_t + \gamma_t(1 - \lambda_t\beta_t)\\ &= \frac{1}{4\lambda_t(1 - \gamma_t\lambda_t)}.\end{aligned}$$

The conditional variance of the liquidity value for the uninformed investors can be shown to be

$$
\begin{aligned}
(4.101) \qquad \Sigma_t &= Var[v|p_t] \\
&= Var[v|\Omega_{t-1}] - \frac{Cov[v, \Delta x_t + u_t|\Omega_{t-1}]^2}{Var[\Delta x_t + u_t|\Omega_{t-1}]} \\
&= (1 - \lambda_t \beta_t)\Sigma_{t-1}.
\end{aligned}
$$

Equations (4.95) - (4.98), (4.100) and (4.101) completely characterize the equilibrium. In order to avoid profits of the informed investor after the last auction, we have to impose the boundary condition $\delta_T = \gamma_T = 0$. It can now be proved that with these two conditions the equilibrium exists, i.e. (4.92) is fulfilled, and that it is a unique linear equilibrium.[28]

As we see from (4.94) $1 - \lambda_t \beta_t$ is always between zero and one, hence the variance of the liquidity value for the uninformed investors is strictly decreasing as long as $0 < \sigma^2_{u^i,t} < \infty$. Therewith the information is gradually incorporated into the price. An efficient market in the strong form is not achieved immediately, i.e. prices do not reveal fully all information available in the market, including the information of the insider. They only tend to become fully revealing over time. This behavior is the result of the profit maximization of the insider. He exploits his informational advantage over time by placing only small orders to hide his trades in the trades of noise traders.

Figure 4.1 illustrates the behavior of the liquidity parameter λ_t depending on the time and the number of auctions.[29] It can be seen that λ_t is decreasing, i.e. liquidity is increasing, with time for all T. But as T increases, λ_t becomes nearly constant over time. It is optimal for the informed investor is to hold λ_t constant, the more auctions there are, the better he can achieve this situation. Hence by placing an order the price is always equally affected, i.e. trading costs of uninformed investors are constant over time.

[28] See KYLE (1985, pp. 1325 f.) for a formal proof. As before there can exist other, nonlinear equilibria.

[29] A method to solve the dynamic equations has been proposed by HOLDEN AND SUBRAHMANYAM (1992, pp. 253 f.).

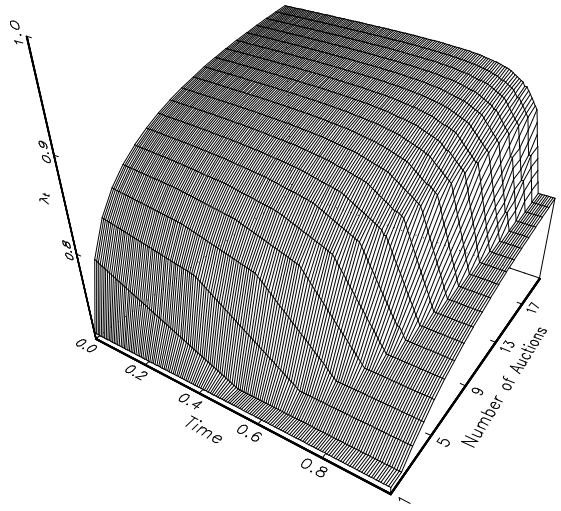


Figure 4.1: Liquidity in the KYLE (1985) model

4.2.2 Auctions with multiple informed investors

HOLDEN AND SUBRAHMANYAM (1992) extended the model of KYLE (1985) to incorporate multiple informed investors. The only change that has to be made in the notation is to view the informed investors order flow, Δx_t, to be composed of the order flows of M informed investors, Δx_t^i. Denoting $\Delta \overline{x}_t^i$ the conjecture of informed investor i about the average order flow of the other informed investors, we have

$$\Delta x_t = \sum_{k=1}^{M} \Delta x_t^k = \Delta x_t^i + (M-1)\Delta \overline{x}_t^i. \tag{4.102}$$

As all informed investors are assumed to be equal, the only reasonable conjecture about the other informed investors' behavior is that they behave exactly in the same way, i.e. $\Delta \overline{x}_t^i = \Delta x_t^i$.

With the other assumptions identical to KYLE (1985), the derivation follows exactly the steps already presented in the last section. Any informed investor maximizes his expected profits given his set of information, $\Omega_p = \{p_1, \ldots, p_t, v\}$, in analogy to the KYLE (1985) model presented above.

Following the same steps as above it is straightforward to show that

$$
\begin{aligned}
\lambda_t &= \frac{M\beta_t\Sigma_{t-1}}{M^2\beta_t^2\Sigma_{t-1}+\sigma^2_{u^i,t}}, \\
\beta_t &= \frac{1-2\gamma_t\lambda_t}{\lambda_t(1+M(1-2\gamma_t\lambda_t))}, \\
\mu_t &= p_{t-1}, \\
\alpha_t &= -\beta_t p_{t-1}, \\
\delta_{t-1} &= \delta_t + \gamma_t\lambda_t\sigma^2_{u^i,t}, \\
\gamma_{t-1} &= \frac{1-\gamma_t\lambda_t}{\lambda_t(1+M(1-2\gamma_t\lambda_t))^2}.
\end{aligned} \tag{4.103}
$$

The conditional variance of the liquidation value for uninformed investors from observing the prices is given by

$$
\begin{aligned}
\Sigma_t &= Var[v|p_t] \\
&= (1-M\lambda_t\beta_t)\Sigma_{t-1}.
\end{aligned} \tag{4.104}
$$

As in KYLE (1985) the information is gradually incorporated into the price because the variance is strictly decreasing over time.

While a single informed investor uses its monopoly power to hold λ_t constant over time, competition between informed investors forces them to trade more aggressively on their information in the first auctions. In consequence, most information will be revealed in these first auctions, resulting in a lower liquidity of the market and a faster decreasing variance. Information is revealed much faster than with a monopolistic informed investor. As nearly all information has been revealed in the first auctions, later trades are not very informative and λ_t quickly approaches zero, i.e. the market becomes very liquid in later auctions. As the number of informed investors increases, information is revealed faster. In the limit as M reaches infinity, prices are fully revealing, i.e. strongly efficient, in an instant with the first trade.

With perfect competition therefore the market is efficient in the strong form, otherwise only in the semistrong form. If competition is too strong, profits from trading on this information are very low and may not cover the costs of acquiring information, although the prices will never be fully revealing as $\Sigma_t > 0$ in all auctions due

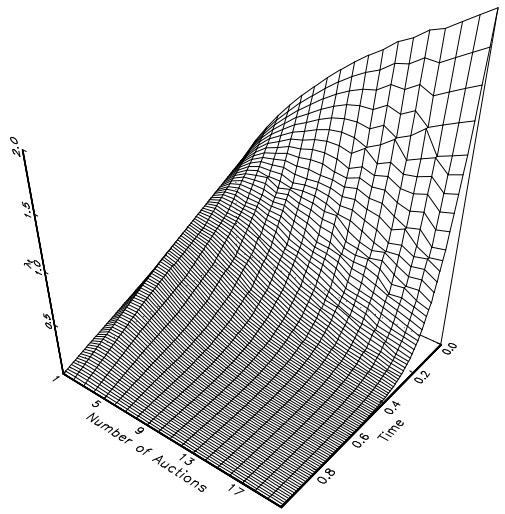


Figure 4.2: Liquidity with different number of auctions

to the noise of uninformed investors. This gives rise to the problem already pointed out by GROSSMAN AND STIGLITZ (1980) that markets cannot be fully revealing if information is costly. If with fully revealing prices profits from information acquisition are too low to cover these costs, no investor will acquire information and prices are not informative. But this on the other hand gives incentives for investors to acquire information and make profits, hence no equilibrium in information acquisition exists.

Figures 4.2 - 4.5 illustrate the findings of this model.

4.2.3 Auctions with risk averse uninformed investors

Thus far we assumed that uninformed investors trade for exogenous reasons and do not respond to price changes, i.e. they were noise traders. We will now assume that they are risk averse investors holding a portfolio consisting of the risky asset and a riskless asset. Suppose that uninformed investors face an exogenously given portfolio imbalance, w_j. This imbalance can be due to changed prices, new information or

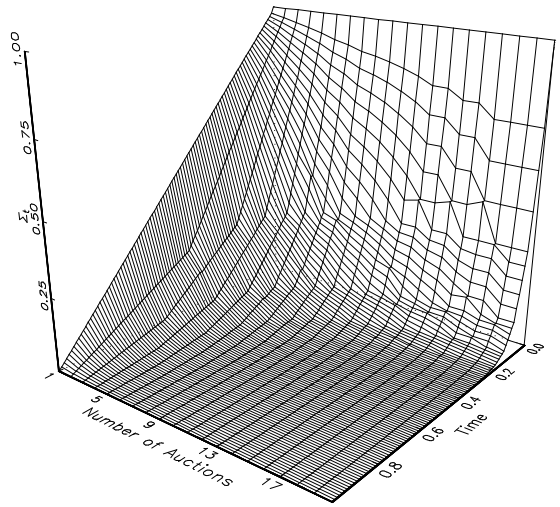


Figure 4.3: Informativeness of prices with different number of auctions

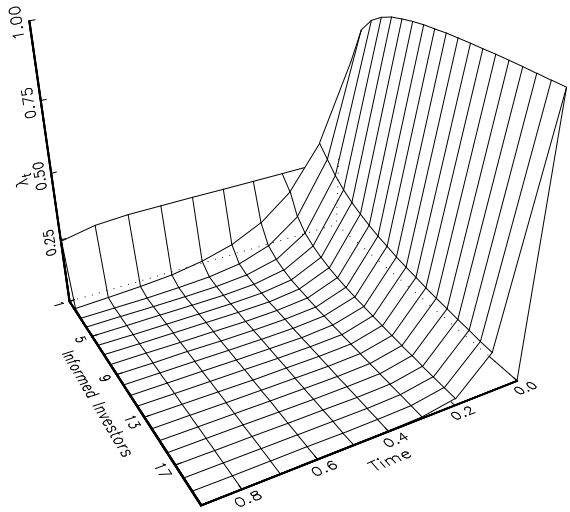


Figure 4.4: Liquidity with different number of informed investors

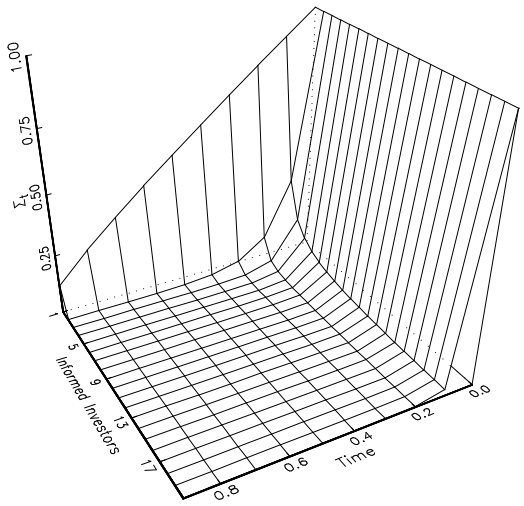


Figure 4.5: Informativeness of prices with different number of informed investors

liquidity needs. They will not only trade this imbalance as assumed before, but act to maximize their expected utility, i.e. they are hedgers.

We assume that the portfolio imbalance is a normally distributed random variable with mean zero and variance $\sigma_w^2 > 0$. The imbalance is assumed to be independent between investors and from any other relevant variable.[30]

$$w_j \sim N(0, \sigma_w^2) \tag{4.105}$$

for all $j = 1, \ldots, N$, where N denotes the number of uninformed investors.

There is only a single auction before the asset is liquidated. Again only linear equilibria are considered, the orders of uninformed investors are also linear in their

[30] The model presented here follows SPIEGEL AND SUBRAHMANYAM (1992). Although they provide a more general framework, we restrict its structure to make the results comparable to the models presented in the previous sections.

portfolio imbalance:

$$u_j = \eta + \xi w_j \quad j = 1, \ldots, N, \tag{4.106}$$

$$x_i = \alpha + \beta v \quad i = 1, \ldots, M, \tag{4.107}$$

$$p = \mu + \lambda(x + u), \tag{4.108}$$

where $x = \sum_{i=1}^{M} x_i$ and $u = \sum_{j=1}^{N} u_j$.

The derivation of the equilibrium follows the same steps as before. The informed investors, assumed to be risk neutral again, maximize their expected profits given their information on the fundamental value.

Using the same methods as above we find that

$$\begin{aligned} \lambda &= \sqrt{\frac{\Sigma_0}{N\sigma_w^2}\frac{M}{(M+1)^2\xi^2}}, \\ \beta &= \sqrt{\frac{\sigma_w^2 N}{\Sigma_0}\frac{\xi^2}{M}}, \\ \alpha &= -\beta p_0, \\ \mu &= p_0 - N\lambda\eta. \end{aligned} \tag{4.109}$$

Uninformed investors maximize their expected utility from holding a portfolio. For simplicity we do not consider the utility derived from holding the optimal portfolio, but only the utility of a deviation from this portfolio. The deviation consists of the size of the portfolio imbalance, w_j and the amount traded, u_j, adjusted by the price paid. The wealth of a deviation from the optimal portfolio is:

$$W_j = v(u_j + w_j) - u_j p \tag{4.110}$$

Inserting the expressions for the parameters and the pricing rule, it is straightforward to show that

$$E[W_j|w_j] = -\lambda u_j^2 + p_0 w_j - N\lambda\eta u_j, \tag{4.111}$$

$$\begin{aligned} Var[W_j|w_j] &= (w_j + v_j)^2\Sigma_0 - 2(u_j + w_j)\lambda\beta u_j M\Sigma_0 \\ &\quad + u_j^2\left(M^2\lambda^2\beta^2\Sigma_0 + (N-1)\lambda^2\xi^2\sigma_w^2\right). \end{aligned} \tag{4.112}$$

The expected utility of uninformed investors is given by

$$E\left[U(W_j)|w_j\right] = U\left(E[W_j|w_j] - \frac{1}{2}zVar[W_j|w_j]\right). \tag{4.113}$$

Maximizing (4.113) after having inserted (4.111) and (4.112) gives the following first order condition:

$$\begin{aligned} 0 &= u_j\left(-2\lambda - z\Sigma_0(1-M\lambda\beta)^2 - z(N-1)\lambda^2\xi^2\sigma_w^2\right) \\ &\quad -zw_j\Sigma_0(1-M\lambda\beta) - N\lambda\eta. \end{aligned} \tag{4.114}$$

The second order condition for a maximum $-2\lambda - z\Sigma_0(1-M\lambda\beta)^2 - z(N-1)\lambda^2\xi^2\sigma_w^2 < 0$ is fulfilled as $\lambda > 0$ from the second order condition of the profit maximization of informed investors and all other terms are positive. Solving for the optimal trade size of uninformed investors yields

$$\begin{aligned} u_j &= -\frac{N\lambda\eta}{2\lambda + z\Sigma_0(1-M\lambda\beta)^2 + z(N-1)\lambda^2\xi^2\sigma_w^2} \\ &\quad -\frac{z\Sigma_0(1-M\lambda\beta)}{2\lambda + z\Sigma_0(1-M\lambda\beta)^2 + z(N-1)\lambda^2\xi^2\sigma_w^2}w_j. \end{aligned} \tag{4.115}$$

By comparing coefficients with (4.106) we see that

$$\begin{aligned} \xi &= -\frac{z\Sigma_0(1-M\lambda\beta)}{2\lambda + z\Sigma_0(1-M\lambda\beta)^2 + z(N-1)\lambda^2\xi^2\sigma_w^2}, \\ \eta &= -\frac{N\lambda\eta}{2\lambda + z\Sigma_0(1-M\lambda\beta)^2 + z(N-1)\lambda^2\xi^2\sigma_w^2}. \end{aligned} \tag{4.116}$$

The last equation implies

$$\eta = 0. \tag{4.117}$$

The first equation has to be solved for ξ:

$$z(N-1)\lambda^2\xi^3\sigma_w^2 + \left(z\Sigma_0(1-M\lambda\beta)^2 + 2\lambda\right)\xi + z\Sigma_0(1-M\lambda\beta) = 0 \tag{4.118}$$

As all coefficients are positive[31] the solution for ξ has to be negative. Inserting for β and λ we find that

$$\xi = -\frac{\left(z\Sigma_0\sqrt{\sigma_w^2 N} - 2\sqrt{\Sigma_0 M}\right)N(M+1)}{\sqrt{\sigma_w^2 N}z\Sigma_0(M(N-1)+N)}. \tag{4.119}$$

[31] It can easily be shown from (4.109) that $1 - M\lambda\beta = 1 - \frac{M}{M+1} = \frac{1}{M+1} > 0$.

As ξ has to be negative, a linear equilibrium only exists if the numerator is positive as the denominator always is positive, i.e. if

$$\begin{aligned} z\Sigma_0\sqrt{\sigma_w^2 N} - 2\sqrt{\Sigma_0 M} &> 0, \\ M &< \frac{1}{4}Nz^2\Sigma_0\sigma_w^2. \end{aligned} \tag{4.120}$$

The existence depends on a not too large fraction of informed investors in the market. The higher the risk aversion, uncertainty about the liquidation value, Σ_0, and dispersion of portfolio imbalances, the more informed investors can participate in the market. This can be explained with the behavior of uninformed investors, as in these cases their need to trade is higher and they are willing to make larger losses to rebalance their portfolios. If this condition is not fulfilled, the losses from trading are too large for uninformed investors such that they leave the market. Therewith informed investors cannot hide their trades, they make no profits and hence do not trade and the market breaks down.

We can write the found equilibrium as

$$p = p_0 + \lambda(x + u), \tag{4.121}$$

$$x_i = \beta(v - p_0), \tag{4.122}$$

$$u_j = \xi w_j, \tag{4.123}$$

where λ, β and ξ are given by (4.109) and (4.119).

The variance of the liquidation value after observing the price is given by

$$\begin{aligned} \Sigma_1 &= Var[v|p] \\ &= \frac{\Sigma_0}{M+1}. \end{aligned} \tag{4.124}$$

The informativeness of the price does only depend on the number of informed investors. Competition between them to trade on their information results in a high revelation of information. The higher trading volume of uninformed investors arising from more risk averse uninformed investors or larger order imbalances, is exactly compensated by more aggressive trading of informed investors.

Analyzing the equilibrium, we find that λ is decreasing in the risk aversion of uninformed investors, hence liquidity increases. If uninformed investors are more risk averse, their wish to trade is larger, although they make losses from trading. This increased trading of uninformed investors enables informed investors to hide their trades better and the order flow becomes less informative. The same holds if the portfolio imbalance increases, i.e. σ_w^2 increases. The need to offset the imbalance at least partially is increased and trades of uniformed investors increase, hence liquidity rises.

If the number of uninformed investors increases, the unconditional volatility of the price increases:

$$
\begin{aligned}
(4.125) \qquad Var[p] &= Var[p_0 + \lambda(x+u)] \\
&= \lambda^2 Var\left[M\beta(v-p_0) + \xi\sum_{j=1}^{N} w_j\right] \\
&= M^2\lambda^2\beta^2\Sigma_0 + \lambda^2\xi^2 N^2\sigma_w^2.
\end{aligned}
$$

As the price of the trade is not known at the time the order is submitted, the risk from trading is increased for uninformed investors. If they are not too risk averse, this effect will be more than compensated by the additional orders from the increased number of uninformed investors, i.e. adverse selection costs are smaller and λ will be decreasing in the number of uninformed investors. If uninformed investors are more risk averse, however, the reduced order from every single uninformed investor dominates the additional order flow generated by more uninformed investors in the market and λ increases with more uninformed investors. But if the number of uninformed investors is sufficiently enlarged, the additional order flow will dominate again and λ decreases. Figure 4.6 illustrates this finding.

By increasing the number of informed investors, competition between them is increased as has been shown in the previous section, resulting in an increased λ. On the other hand, as we saw in (4.124), the informativeness of prices is increased, i.e. the adverse selection costs of the uniformed investors are reduced. If the uninformed investors are very risk averse this will induce them to trade more actively,

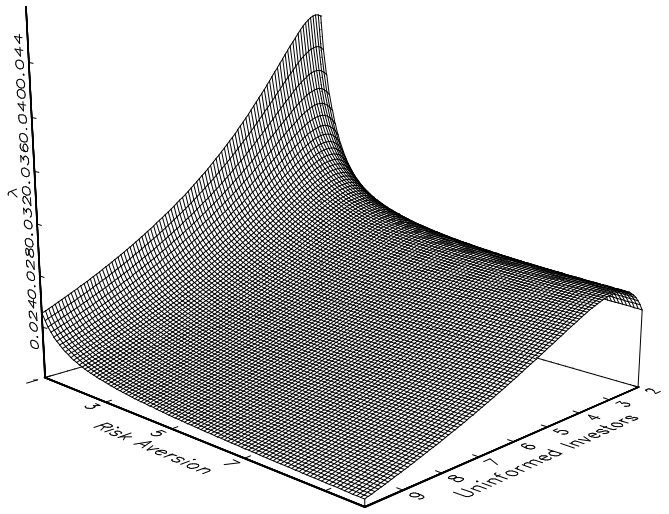


Figure 4.6: Market liquidity with a varying number of uninformed investors

thereby compensating the increased λ, which becomes decreasing. If they are less risk averse their increased trading cannot compensate the effect of competition and λ is increasing as illustrated in figure 4.7.

If the uncertainty about the liquidation value, Σ_0, increases, the adverse selection costs and the variance of the price will increase, hence uninformed investors will scale back their trades, this increases λ. On the other hand the benefits from offsetting an imbalance is increased. If uninformed investors are very risk averse, the first effect dominates and λ will be increasing in Σ_0. If the uninformed investors are less risk averse, the second effect will dominate for small uncertainties and λ decreases, if the uncertainty is too large, however, the first effect again will dominate and λ increases again. This behavior is illustrated in figure 4.8.[32]

Spiegel and Subrahmanyam (1992) provide a more general framework by allowing the signals the informed investors receive to be imperfect and differ between investors. Compared to the more restrictive version presented here, no additional

[32] Formal proofs of these claims are presented in Spiegel and Subrahmanyam (1992).

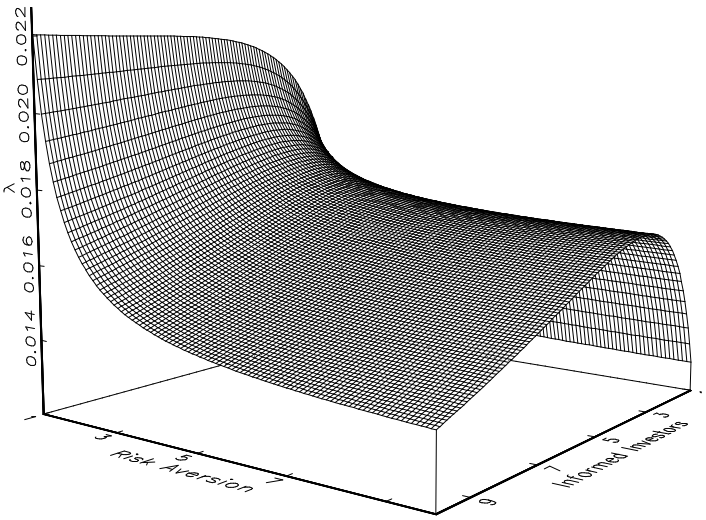


Figure 4.7: Market liquidity with a varying number of informed investors

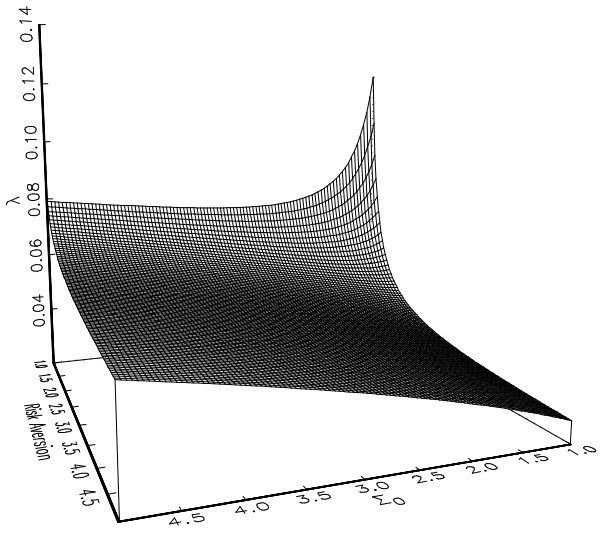


Figure 4.8: Market liquidity with different uncertainties about the liquidation value

insights can be gained, whereas the tractability of the derivation is reduced significantly. Introducing risk averse informed investors does also not change the results derived here, but becomes in a general framework not solvable analytically.[33]

A central element for calculating conditional moments and hence deriving the equilibrium is the assumption of normally distributed random variables. This assumption cannot be lifted without facing the problem that the results may change. FOSTER AND VISWANATHAN (1993a) derive similar results within the more general class of elliptically contoured distributions instead of using the special case of a normal distribution, a generalization to other distributions cannot be made.

All models presented here, considered only linear equilibria, nothing can be said about the existence and properties of non-linear equilibria. KYLE (1985, p. 1322) suspects that there exist no non-linear equilibria, but he is not able to proof his suspicion.

There exist several extensions of these models, which we will not consider in detail. BONDARENKO (1999) relaxes the assumption of only a single match maker by investigating a model with a given number of match makers that are not restricted to quote prices such that they make zero expected profits, instead they apply strategic price setting to maximize their profits. He finds the results as derived above to hold with the number of match makers reaching infinity, i.e. if they behave competitively. Strategic behavior of informed investors is investigated in KYLE (1989). He finds that the prices do not tend to become strongly efficient in the long run and with an increasing number of informed investors the competitive outcome is not achieved.

BACK ET AL. (2000) consider insiders having diverse information and show that prior to the last trading round, the market becomes completely illiquid with a substantial amount of private information not revealed. They further show that in late

[33] SUBRAHMANYAM (1991) presents a model with risk averse informed investors. He uses only noise traders, but no hedgers in his model. The results he derives are very similar to those with risk neutral informed investors, so that no new insights are gained.

trading rounds the informational efficiency with a monopolistic insider would have been higher.

By adding another source of information on the value of the asset for the match maker, JAIN AND MIRMAN (1999) show that he sets more informative prices and reduces the profits of informed investors. CHAU (1998) considers the trading of many pure noise traders, a single risk averse informed investor and a risk averse market maker. Although this model assumes a dealer rather than an auction market, it helps understanding the adjustment of prices to new information. In contrast to models with risk neutral market participants, he finds that risk averse informed investors can cause prices to overshoot before they gradually adjust to the new fundamental value. He attributes this behavior to portfolio imbalances building up for the informed investor when trading upon information. After having exploited most of the information, he faces a large portfolio imbalance, which he tries to reduce in the following. This behavior causes price changes to reverse and the new fundamental value is approached gradually after a first overshooting.

4.2.4 The informational content of trading volume

Thus far only prices have been used by uninformed investors to revise their beliefs on the fundamental value of the asset. Another variable is widely used in markets as a source of information: trading volume. Using this additional information enables uninformed investors to learn the information more quickly. BLUME ET AL. (1994) provide a simple model of trading volume used by uninformed investors to deduct additional information.

They assume the fundamental value to be normally distributed with mean p_0 and variance σ^2:

$$p \sim N(p_0, \sigma^2). \tag{4.126}$$

A share of γ of the total N investors are assumed to be informed, but unlike in the previous models their information is not perfect. They only observe a noisy signal

of the fundamental value:[34]

$$\psi_1^i = p^* + \varepsilon_1^i, \tag{4.127}$$

where

$$p^* \sim N(p, \sigma_p^2) \tag{4.128}$$

is a common signal. There remains uncertainty about the transformation of the information into the fundamental value. This information is not observed purely, i.e. there is some noise with

$$\varepsilon_1^i \sim N(0, \sigma_{\varepsilon_1}^2). \tag{4.129}$$

The error variance $\sigma_{\varepsilon_1}^2$ is itself a random variable, i.e. the quality of information varies randomly over time. Informed investors know the realization of $\sigma_{\varepsilon_1}^2$ when they receive their information, uninformed investors only know its distribution. Uninformed investors are assumed to receive also the signal p^* but with another error term ε_2:

$$\begin{aligned} \psi_2^i &= p^* + \varepsilon_2^i, \\ \varepsilon_1^i &\sim N(0, \sigma_{\varepsilon_2}^2), \end{aligned} \tag{4.130}$$

where $\sigma_{\varepsilon_2}^2$ is known to all investors, informed and uninformed. With the assumption that $\sigma_{\varepsilon_2}^2 > \sigma_{\varepsilon_1}^2$ the information of informed investors is more precise, that is why they are called informed. We therewith have for the distribution of the true value ψ_i of group i, if we assume the errors terms to be independent of all other relevant variables:

$$\psi_i \sim N(p, \sigma_p^2 + \sigma_{\varepsilon_i}^2). \tag{4.131}$$

The demand of the investors for the asset in this setting has been derived by Diamond and Verrecchia (1981) and Brown and Jennings (1989). When we restrict the equilibrium to be linear we receive with risk averse investors j the demand

[34] This modification is necessary as otherwise uninformed investors would be able to deduct the information completely by only observing price and volume. In this case informed investors would not be able to make profits from their information and if information is costly there would be no incentives to become informed. Hence, as pointed out by Grossman and Stiglitz (1980) no equilibrium in information acquisition would exist.

for every individual in groups 1 (informed investors) and 2 (uninformed investors) as

$$
\begin{aligned}
d_1^j &= \frac{p_0 - p_1}{z\sigma^2} + \frac{\psi_1^j - p_1}{z(\sigma_p^2 + \sigma_{\varepsilon_1}^2)}, \\
d_2^j &= \frac{p_0 - p_1}{z\sigma^2} + \frac{\psi_2^j - p_1}{z(\sigma_p^2 + \sigma_{\varepsilon_2}^2)},
\end{aligned} \tag{4.132}
$$

where z denotes the Arrow-Pratt measure of risk aversion and p_1 the price applied.[35] The total demand of all investors has to equal zero in equilibrium as long as the amount of the asset is fixed. Aggregation over all individuals finally gives us after dividing by N and multiplying with the risk aversion z:

$$
0 = \frac{p_0 - p_1}{\sigma^2} + \frac{\gamma(\overline{\psi}_1 - p_1)}{\sigma_p^2 + \sigma_{\varepsilon_1}^2} + \frac{(1-\gamma)(\overline{\psi}_2 - p_1)}{\sigma_p^2 + \sigma_{\varepsilon_2}^2}, \tag{4.133}
$$

where $\overline{\psi}_1 = \frac{1}{N}\sum_{i=1}^{\gamma N} \psi_1^i$ and $\overline{\psi}_2 = \frac{1}{N}\sum_{i=\gamma N+1}^{N} \psi_2^i$ denote the average realization of the information. Solving for p_1 gives

$$
p_1 = \left[\frac{p_0}{\sigma^2} + \frac{\gamma}{\sigma_p^2 + \sigma_{\varepsilon_1}^2}\overline{\psi}_1 + \frac{1-\gamma}{\sigma_p^2 + \sigma_{\varepsilon_2}^2}\overline{\psi}_2\right]\left[\frac{1}{\sigma^2} + \frac{\gamma}{\sigma_p^2 + \sigma_{\varepsilon_1}^2} + \frac{1-\gamma}{\sigma_p^2 + \sigma_{\varepsilon_2}^2}\right]^{-1}. \tag{4.134}
$$

If we let $N \to \infty$ then the law of large numbers can be used to show that $\overline{\psi}_j \to p$ and (4.134) becomes

$$
p_1 = \left[\frac{p_0}{\sigma^2} + \left(\frac{\gamma}{\sigma_p^2 + \sigma_{\varepsilon_1}^2} + \frac{1-\gamma}{\sigma_p^2 + \sigma_{\varepsilon_2}^2}\right)p\right]\left[\frac{1}{\sigma^2} + \frac{\gamma}{\sigma_p^2 + \sigma_{\varepsilon_1}^2} + \frac{1-\gamma}{\sigma_p^2 + \sigma_{\varepsilon_2}^2}\right]^{-1}. \tag{4.135}
$$

The total trading volume is the absolute value of the demands given in (4.132). To avoid double counting this has to be divided by two. Hence the trading volume, adjusted to per capita average volume, is with $N \to \infty$:

$$
\begin{aligned}
V &= \frac{1}{2}\frac{1}{N}\left(\sum_{i=1}^{\gamma N} |d_1^i| + \sum_{i=\gamma N+1}^{N} |d_2^i|\right) \\
&= \frac{\gamma}{2}\frac{1}{\gamma N}\sum_{i=1}^{\gamma N} |d_1^i| + \frac{1-\gamma}{2}\frac{1}{(1-\gamma)N}\sum_{i=\gamma N+1}^{N} |d_2^i| \\
&\to \frac{\gamma}{2}E[|d_1|] + \frac{1-\gamma}{2}E[|d_2|].
\end{aligned} \tag{4.136}
$$

[35] We derived a similar result in chapter 4.1.1 as equation (4.23). The derivation of the above formula follows similar considerations.

Inserting from (4.132) and (4.135) gives the expression for the volume, which is explicitly stated in BLUME ET AL. (1994, p. 165), but not reproduced here. The evaluation is best conducted with numerical examples.[36]

If the information received by informed investors is only of low precision and suggests only a small deviation from prior beliefs, i.e. $\sigma^2_{\varepsilon_1}$ is relative large compared to $\sigma^2_{\varepsilon_2}$, informed investors have not much confidence in their information. Although their beliefs are widely spread, they do not trade much on their information, hence trading volume will be relatively small. If the precision increases, i.e. $\sigma^2_{\varepsilon_1}$ decreases, they become more confident in their information and if their beliefs are wide enough dispersed, trading volume increases. For $\sigma^2_{\varepsilon_1} = \sigma^2_p$ the trading volume reaches its maximum. If the precision of information is further increased, their confidence also increases, but on the other hand the dispersion of beliefs is reduced, they do not find many informed investors to trade with and are forced to trade nearly only with uninformed investors, hence trading volume reduces.

If the information received suggests a large deviation from the prior belief of p_0, the rebalancing of the portfolios will dominate, even if the precision is not high. The trading volume will increase with the precision of the information, even if the dispersion of beliefs is reduced.

The larger the signal suggests the deviation from prior beliefs is, the more need the investors to rebalance their portfolios, this need increases with the precision of information. As p_1 denotes also the new belief of the informed investors, $p_1 - p_0$ denotes the change in belief. Therewith the volume should be V-shaped in the change of beliefs, where the V becomes more pronounced the more precise the information is. The lowest trading volume should be found at $p_0 = p_1$.

Figure 4.9 visualizes these findings using the expression for trading volume provided by BLUME ET AL. (1994). Virtually all empirical investigations on price changes

[36] The results of the illustrations below can be shown to be valid in all economically relevant situations.

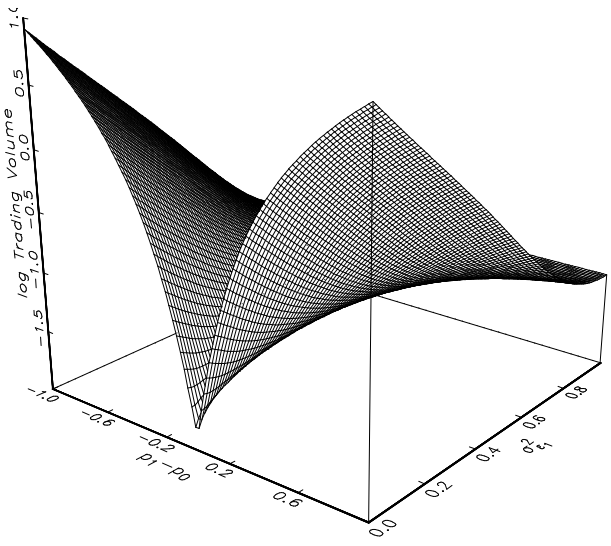


Figure 4.9: Trading volume with different precision of information and changes of beliefs

and volume find V-shaped pattern between the price change and volume. With this simple model this pattern can be explained.

Figure 4.10 shows furthermore that the V-shape does not only depend on the precision of information, but also on the dissipation of information, i.e. the share of informed investors γ. The more investors are informed the less pronounced the V-shape is, until it vanishes if all investors are informed. This pattern can be explained with the fact that with a price change due to new information only few investors can expect to make profits as their beliefs deviate from the current price. Hence the trading volume does not respond so sensitive to price changes. With only minor price changes, or equivalently changes in beliefs, the volume is increasing for $\sigma^2_{\varepsilon_1} > \sigma^2_p$ and decreasing for $\sigma^2_{\varepsilon_1} < \sigma^2_p$ due to the effect of rebalancing portfolios as described above.

While prices reveal information about the magnitude of a signal, i.e. the change in beliefs of the informed investors, trading volume reveals information about the precision and dissemination of information. By observing prices and trading volume,

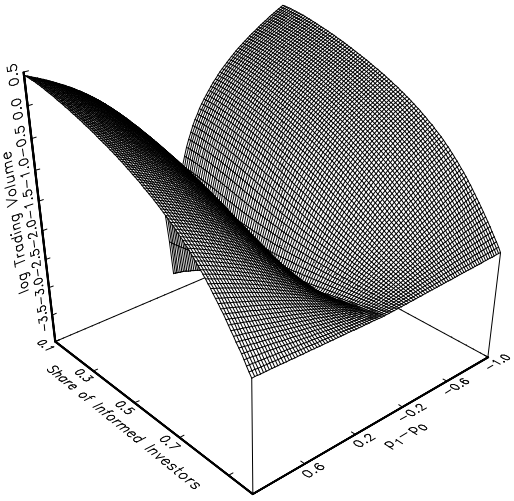


Figure 4.10: Trading volume with different shares of informed investors and changes of beliefs

an investor can find out the beliefs of the informed investors and their changes by observing prices and the persistence of the movement, i.e. the confidence of informed investors by observing trading volume. This can be observed in markets where large price changes in association with low trading volume in most cases are viewed as being not very persistent, whereas the same price change with a large trading volume is viewed as persistent by market observers.

4.2.5 Explaining short-term movements of asset prices

In recent years many empirical studies have been conducted to detect patterns in asset markets. Many such patterns have been found between trading days as well as within a trading day. Various explanations and models have been offered to explain the found behavior. In this section we will not cover this literature in detail, but concentrate on two widely cited contributions, whose results can be explained intuitively with the models presented above, such that there is no need to develop the models actually chosen by the authors in detail.

FOSTER AND VISWANATHAN (1990) use the above derived framework to explain the variation of asset price volatility and trading volume within a trading week. Trading takes place typically from Monday to Friday, while they assume that information is acquired on every day of the week, i.e. from Monday to Sunday, but portfolio imbalances are not aggregated over the weekend. Therewith informed investors should have a larger informational advantage from trading on Mondays, on the other hand adverse selection costs are highest on Mondays. Hence uninformed investors will not trade so actively on this day. If we assume further that a part of the information becomes public by other sources than prices, e.g. publications in newspapers or rumors, over the following trading days, informed investors are forced trade on their information on Mondays in order to make profits, hence the share of informed investors will be relative high on these days and trading volume should be low due to the reduced participation of uninformed investors. A high share of informed investors implies a large change in the price due to the reduced liquidity of the market, hence we should expect to find a larger volatility on Mondays. Empirical investigations support this effect as suggested by this model.

Due to adverse selection costs, uninformed investors always make losses when trading with informed investors. The more uninformed investors are in the market, the more these costs are distributed among them, reducing costs of trading for a single uninformed investor. It would therefore be a beneficial strategy for uninformed investors to concentrate their trading in a certain trading round. Also informed investors would be better able to hide their trades and would be able to trade more and hence make more profits in these trading rounds. It is reasonable to assume that for most investors it is of no importance at which time of the day they trade as the settlement of the orders depends only on the day of trading and not on the time of trading within a day. ADMATI AND PFLEIDERER (1988) show that with such assumptions trading will be concentrated at certain points of time in the trading day, provided competition between informed investors is large enough and adverse selection costs do not increase too much by the presence of more informed investors. The presence of many informed and uninformed investors increases the volatility of

prices in those times of higher trading activity, a result that is confirmed in empirical studies.

FOSTER AND VISWANATHAN (1993b) conduct an empirical investigation of assets listed on the NYSE and find that their results are consistent with the above described behavior.

The two models shortly presented above can explain some of the effects observed in asset markets regarding volatility and trading volume. However, they cannot explain observed patterns in returns. ADMATI AND PFLEIDERER (1989) provide a model how to explain such patterns, but they have to introduce another rule how the match maker determines prices. They assume the match maker to set prices strategically, what leaves the framework of the models presented above.

4.3 Inventory-based models of market making

The last section analyzed how prices adjust to new information, therefore the order flow has been aggregated over a given period of time. The match maker had no other role than to determine the equilibrium price and played no active role in the trading process. In the following sections we will now investigate the behavior of prices on a trade-by-trade basis. We therefore introduce a market maker into the trading process, replacing the match maker. The market maker directly influences the prices by quoting prices at which he is willing to buy and sell the asset. But in general he also has no active role in the trading process as he does not initiate or actively search for a trade, but waits for an order to arrive and then clears this order at the stated price on his own account. It is intuitively clear that transaction prices are not only influenced by the orders, but also by the behavior of the market maker.

There have evolved two main groups of theories modeling the behavior of market makers, inventory and information-based models. Information-based models are

close to the models of auction markets, they assume a risk-neutral market maker and two groups of investors, informed and uninformed investors. Inventory-based models assume a risk averse market maker, all investors have the same information and agree on the implication of this information on the fundamental value, i.e. they agree on the fundamental value. While in information-based models trades can be motivated by exploiting informational advantages (informed investors) and need for liquidity (uninformed investors or noise traders), in inventory-based models need for liquidity is the only source of trade. This section analyzes inventory-based models and the next section information-based models.

As has previously been pointed out in section 4.1, a market maker provides the service of enabling an investor to trade immediately at a given price by forming the counterpart. He then waits for another order offsetting his position. Therewith he takes the risk of not knowing when and at which price he can offset this position. A trade that is typically conducted between two investors is divided into two parts that occur at different prices at different times. Bearing these risks imposes costs on the market maker he has to cover by quoting different prices at which he willing to buy (bid price) and to sell (ask price) the asset. As the investor has not to bear these risks, he is willing to pay for this service by accepting a less favorable price for the trade until the costs imposed by the market maker equals his costs from waiting for an offsetting order and trading at an unknown price (waiting costs).[37]

We now have to make some assumptions on the determination of the order flow.[38] At first we assume that all investors place their orders independently of each other. Each investor either submits a buy or a sell order, but not both. Therewith the order submissions to buy and to sell are independent of each other. The need for trading is exogenously given by a liquidity event, which determines the waiting costs of an investor. This liquidity event is assumed to be a random variable that is independent and identical distributed between investors, hence waiting costs are

[37] Demsetz (1968, p. 37) was the first to introduce the concept of waiting costs into literature. His contribution lead the way to the following literature on market microstructure theory.

[38] These assumptions have first explicitly been stated by Garman (1976, pp. 258 ff.).

a random variable. If the waiting costs are higher than the costs imposed by the market maker,[39] he will submit his order to the market maker, otherwise he will wait for an offsetting order by himself. The higher the costs for trading with the market maker, the less orders are submitted. We can now aggregate all orders (separately for buy and sell orders) and see that the orders submitted in a given period of time follow a Poisson process or, equivalently, that the probability of an order arriving in a certain period of time is Poisson distributed.[40]

A Poisson distribution is characterized by the order arrival rate λ. Let λ_a denote the order arrival rate for buy orders (trades at the ask) and λ_b for sell orders (trades at the bid). Let further p^* denote the fundamental value all investors agree on, p_a the ask and p_b the bid price. The costs the market maker imposes on the investors are $p_a - p^*$ and $p^* - p_b$. Therewith we can reasonably assume that

$$\begin{aligned} \frac{\partial \lambda_b}{\partial p_b} > 0, \qquad & \frac{\partial \lambda_a}{\partial p_a} < 0, \\ \frac{\partial \lambda_b}{\partial p_a} = 0, \qquad & \frac{\partial \lambda_a}{\partial p_b} = 0. \end{aligned} \tag{4.137}$$

Figure 4.11 visualizes these findings. At (p',λ') we have a stochastic equilibrium of individual order arrival rates.[41] However by choosing $p_a = p_b = p'$ the market maker would not be able to cover his costs. By quoting different prices, $p_a > p_b$ he can make a profit as he is ready to buy the asset at a less favorable price than he is willing to sell the asset. If the market maker chooses those two prices as marked in figure 4.11, we observe that $\lambda_b > \lambda_a$. This means it is more likely that a sell order arrives next in the market, i.e. the market maker expects to increase his position in the asset. Similarly he will choose $\lambda_b < \lambda_a$ to decrease his position in the asset and $\lambda_a = \lambda_b$ if he does not want his position to change.

Before determining the costs of market making it is necessary to characterize the trading process and the behavior of the market maker in more detail. We assume

[39] The costs are the difference between the value of the asset and the quoted price.

[40] To see this, it has to be noted that submitting an order or not is a binomial variable as the possible values are to submit or not to submit. By aggregating binomial variables they converge to a Poisson process, see e.g. GREENE (1997, p. 72 f.).

[41] The equilibrium is stochastic as only expected demand and supply equal by having the same order arrival rates. The realized demand and supply may not equal.

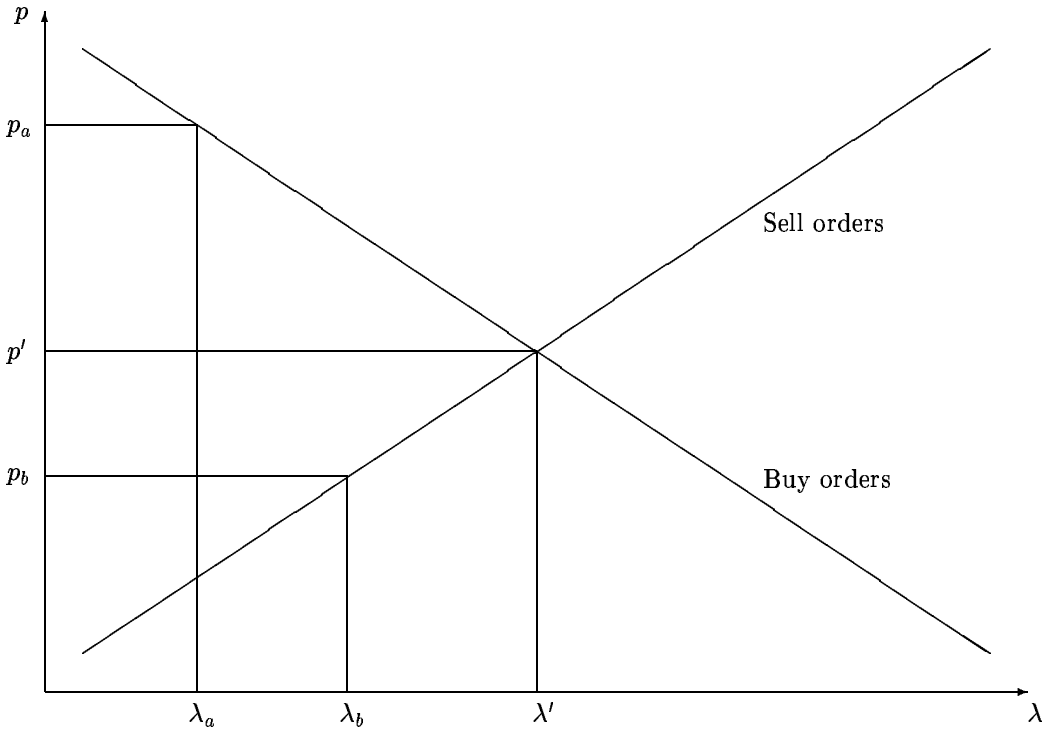


Figure 4.11: Demand and supply in dealer markets

that only a single trade per period of time is submitted to the market maker, either a buy or a sell order.[42] When quoting prices at which he is willing to buy and sell the asset, he does not know whether the next order will be a buy or a sell order.

We consider an economy with only a single risky asset that is traded with the market maker and a riskless asset bearing no interest, e.g. money. Assume further that the market maker has an optimal portfolio consisting of the single risky asset and the riskless asset, chosen according to portfolio selection theory.[43] Any deviations from this optimal portfolio are denoted as *inventory.*[44] After a single trading round we assume the risky asset to be liquidated at the fundamental value.

The market maker is assumed to be risk averse and maximize his expected utility

[42] We could also allow for many orders to be submitted and concentrate on the order imbalance receiving the same results. In this sense we can interpret this situation as an auction market presented above, where the order imbalance has not been served and now is executed by a market maker.

[43] A brief overview of modern portfolio theory is presented in Appendix B.

[44] As the optimal portfolio is fixed we can concentrate our analysis on the inventory that is a linear transformation of the entire portfolio held.

of terminal wealth that occurs after a single round of trading by setting optimal bid and ask prices.

4.3.1 The costs of market making

The first to provide a model how to determine the costs of a market maker was STOLL (1978). Suppose that the market maker holds his optimal portfolio, denoted E in figure 4.12. By accepting a trade, the portfolio actually held deviates from his optimal portfolio, suppose it is located at E', i.e. a buy order arrives on the market and the share of the asset in the portfolio is reduced, while the share of the riskless asset is increased.[45] The utility level from holding portfolio E' instead of the optimal portfolio E decreases from U_0 to U_1. This difference in utility are the costs that the market maker faces from his service. He has to be compensated for this loss in utility, which is done by holding a larger portfolio, i.e. a higher total wealth. In other words, for holding a non-optimal portfolio he is compensated by an increase in the level of holdings.

It is also obvious from figure 4.12 that, if the market maker does currently not hold his optimal portfolio, i.e. his inventory is nonzero, he may gain in utility from accepting an order. Suppose as an example that he holds portfolio E', by accepting a sell order of the same size as before the buy order, he would gain utility by reaching portfolio E again, hence his costs would be negative. In this case p_b would lie above p^* in figure 4.11.

We assume that market makers face only these costs, which are also called *inventory costs*. Other costs, e.g. for order processing, are not included here, but they can easily incorporated into this framework without changing arguments significantly.[46] Early contributions to market microstructure theory were concerned with the ability

[45] As wee know from portfolio theory, the opportunity locus reduces to a straight line if a riskless asset is present.

[46] See STOLL (1978, pp. 1144 ff.). He also provides a more general framework by assuming more than a single risky asset to be in the market. But as he also assumes the market maker to act as market maker only for a single asset the results obtained are identical to those in this restricted version, they only come with more notational burdens.

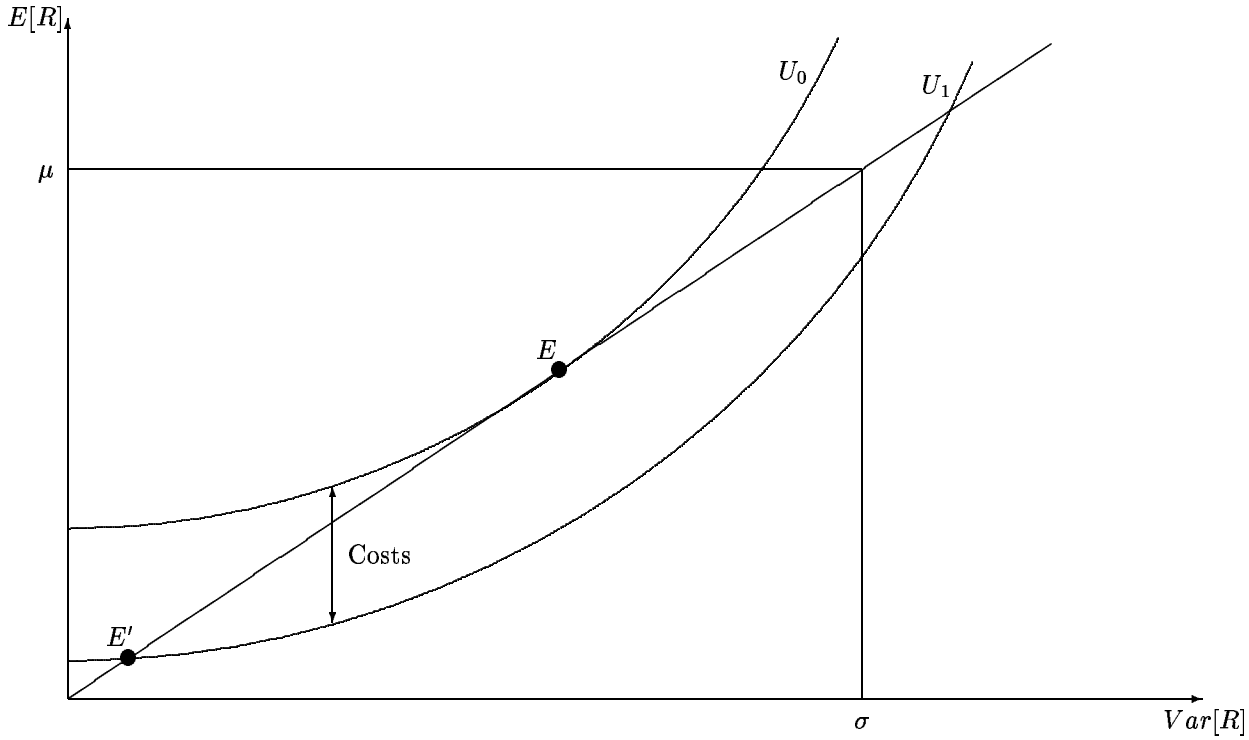


Figure 4.12: Inventory costs of a market maker

of the market maker to deliver the asset or the money. By allowing no short sales the market maker faces the risk of running out of stock, of assets as well as of money.[47] Although such a situation cannot be ruled out by any model, it has been found to be not relevant in practice. By allowing short sales a bankruptcy of the market maker can be avoided and these concerns have not to be considered. As the models focusing on the threat of bankruptcy also give similar results, models of inventory costs have attracted more attention in the literature.

For bearing the above described inventory costs, the market maker has to be compensated, i.e. his expected utility of terminal wealth from holding the initial portfolio and from holding the new portfolio after having accepted a trade, have to be equal. Let W^* denote the terminal wealth without a trade and W the terminal wealth after having accepted a trade. Then the costs are determined such that

$$E\left[U(W^*)\right] = E\left[U(W)\right]. \tag{4.138}$$

[47] See e.g. GARMAN (1976) or AMIHUD AND MENDELSON (1980).

The initial portfolio has not necessarily to be the optimal portfolio, it is the optimal portfolio plus the inventory of the market maker, which can be either positive or negative. Let k denote the fraction of total wealth that is invested into the risky asset in the optimal portfolio, the optimal holding of the risky asset then is kW_0, where W_0 denotes the initial total wealth. The total amount of the risky asset actually held is $kW_0 + I$, with I denoting the inventory. With R denoting the random return of the fundamental value of the risky asset, the final wealth of the portfolio is

$$\begin{aligned} W^* &= W_0 + (kW_0 + I)R \\ &= W_0 \left(1 + \left(k + \frac{I}{W_0}\right) R\right). \end{aligned} \tag{4.139}$$

With $\mu = E[R]$ and $\sigma^2 = Var[R]$ we get

$$E[W^*] = W_0 \left(1 + \left(k + \frac{I}{W_0}\right) \mu\right), \tag{4.140}$$

$$Var[W^*] = W_0^2 \left(k + \frac{I}{W_0}\right)^2 \sigma^2. \tag{4.141}$$

Approximating $U(W^*)$ by a second order Taylor series around $E[W^*]$ we obtain

$$E\left[U(W^*)\right] = U\left(E[W^*] + \frac{1}{2}U''(E[W^*])Var[W^*]\right). \tag{4.142}$$

Let Q denote the trade size, measured in value, not in numbers of assets traded, where $Q > 0$ for a sell order and $Q < 0$ for a buy order. We further denote C as the costs of the market maker to conduct a trade of size Q, transformed from utility into value. We then have for the terminal wealth with accepting a trade:

$$\begin{aligned} W &= W_0 + (kW_0 + I)R + Q(1 + R) - (Q - C) \\ &= W_0 \left(1 + \left(k + \frac{I}{W_0}\right) R\right) + Q(1 + R) - (Q - C), \end{aligned} \tag{4.143}$$

where the first term denotes the part of terminal wealth that arises from holding the initial portfolio, the second term the part that has been affected by the change in inventory due to accepting an order of size Q and the last term the benefits in money from conducting this trade.

From (4.143) we get with assuming the order size Q to be known

$$E[W] = W_0\left(1+\left(k+\frac{I}{W_0}\right)\mu\right)+Q(1+\mu)-(Q-C) \tag{4.144}$$
$$= W_0\left(1+\left(k+\frac{I}{W_0}\right)\mu\right)+Q\mu+C,$$

$$Var[W] = W_0^2\left(k+\frac{I}{W_0}\right)^2\sigma^2+Q^2\sigma^2+2W_0\left(k+\frac{I}{W_0}\right)Q\sigma^2 \tag{4.145}$$
$$= \sigma^2\left(W_0\left(k+\frac{I}{W_0}\right)+Q\right)^2.$$

Approximating $U(W)$ also by a second order Taylor series around $E[W]$ gives

$$E\left[U(W)\right] = U\left(E[W]+\frac{1}{2}U''(E[W])Var[W]\right). \tag{4.146}$$

If Q is relatively small compared to the initial wealth of the market maker we can reasonably assume that

$$U'(E[W^*]) = U'(E[W]), \tag{4.147}$$

$$U''(E[W^*]) = U''(E[W]). \tag{4.148}$$

The mean-value theorem then states that there exists a W' between W and W^* such that

$$\frac{U(E[W])-U(E[W^*])}{E[W]-E[W^*]} = U'(E[W']).$$

With (4.147) we can rewrite this as

$$\frac{U(E[W])-U(E[W^*])}{U'(E[W^*])} = E[W]-E[W^*]. \tag{4.149}$$

By inserting (4.142) and (4.146) into (4.138) we get with (4.148) after rearranging

$$U(E[W])-U(E[W^*]) = \frac{1}{2}U''(E[W^*])(Var[W^*]-Var[W]).$$

With z as the Arrow-Pratt measure of absolute risk aversion[48] we get after dividing by $U'(E[W^*])$ and inserting (4.140), (4.141), (4.144), (4.146) and (4.149):

$$Q\mu+C = \frac{1}{2}z\sigma^2\left(Q^2+2QW_0\left(k+\frac{I}{W_0}\right)\right). \tag{4.150}$$

[48] See appendix A.2 for an introduction of this measure of risk aversion.

For the optimal portfolio, i.e. $I = 0$, we see from (4.139) that $W^* = W_0(1 + kR)$, hence

$$E[U(W^*)] = U\left(E[W^*] - \frac{1}{2}zVar[W^*]\right) \tag{4.151}$$
$$= U\left(W_0(1 + k\mu) - \frac{1}{2}zW_0^2k^2\sigma^2\right)$$

such that

$$\frac{\partial U}{\partial \mu} = W_0kU'(\cdot), \tag{4.152}$$

$$\frac{\partial U}{\partial \sigma^2} = -zW_0^2k^2U'(\cdot). \tag{4.153}$$

By totally differentiating (4.151) we get

$$dE[U(W^*)] = \frac{\partial U}{\partial \mu}d\mu + \frac{\partial U}{\partial \sigma^2}d\sigma^2 \tag{4.154}$$
$$= (W_0kd\mu - zW_0^2k^2d\sigma^2)U'(\cdot).$$

Setting (4.154) equal to zero we get the slope of the indifference curve at the optimal portfolio:

$$\frac{d\mu}{d\sigma^2} = zW_0k. \tag{4.155}$$

The slope of the opportunity locus is known from portfolio selection theory to be

$$\frac{d\mu}{d\sigma^2} = \frac{\mu}{\sigma^2}. \tag{4.156}$$

From portfolio selection theory we know that these two slopes have to be identical, hence we get from these two relations after rearranging:

$$k = \frac{\mu}{zW_0\sigma^2}. \tag{4.157}$$

Inserting (4.157) into (4.150) we get

$$C = -Q\mu + \frac{1}{2}z\sigma^2\left(Q^2 + 2QW_0\left(\frac{\mu}{zW_0\sigma^2} + \frac{I}{W_0}\right)\right) \tag{4.158}$$
$$= -Q\mu + \frac{1}{2}z\sigma^2Q^2 + Q\mu + z\sigma^2QI$$
$$= z\sigma^2\left(\frac{1}{2}Q^2 + QI\right).$$

This is the expression for the costs a market maker faces when accepting orders of size Q. The costs depend on a characteristic of the asset, the variance of the fundamental value σ^2, a characteristic of the trade, the trade size Q, and two characteristics of the market maker, his risk aversion z and his inventory position I.

It is more natural to assume the trade size to be always positive. Define $Q' = |Q|$ as the trade size and we get the costs for a trade at the ask, C_a, and at the bid, C_b, by

$$C_b = z\sigma^2\left(\frac{1}{2}Q^2 + Q'I\right), \tag{4.159}$$

$$C_a = z\sigma^2\left(\frac{1}{2}Q^2 - Q'I\right). \tag{4.160}$$

The relative costs are given by

$$c_b = \frac{C_b}{Q'} = z\sigma^2\left(\frac{1}{2}Q' + I\right), \tag{4.161}$$

$$c_a = \frac{C_a}{Q'} = z\sigma^2\left(\frac{1}{2}Q' - I\right). \tag{4.162}$$

When buying the asset, the market maker reduces the price at which he is willing to buy compared to the fundamental value and increases it when selling the asset. Therefore the *reservation prices* of a market maker for the entire trade of size Q' are

$$p_b = p^* - C_b, \tag{4.163}$$

$$p_a = p^* + C_a. \tag{4.164}$$

The difference between the bid and the ask price is denoted the *spread*, s. From (4.159) - (4.164) we get

$$s = p_a - p_b = C_a + C_b = z\sigma^2 Q^2. \tag{4.165}$$

The spread does not depend on the inventory of the market maker, but only on his risk aversion, the trade size, and the variance of the fundamental value of the asset.[49] Unless the market maker is risk neutral, i.e. $z = 0$ or the asset is riskless,

[49] The reservation prices for trading a single unit of the asset, as usually published, are given by

$$p_b = p^*(1 - c_b),$$
$$p_a = p^*(1 + c_a).$$

i.e. $\sigma^2 = 0$, the spread will always be positive.[50] We therewith have verified the result stated earlier by arguments of arbitrage that

$$p_a > p_b. \tag{4.166}$$

The inventory does not influence the spread, but only the level of prices. It can easily be verified that if $I \neq 0$ the spread is not located symmetrically around p^*. If the inventory is sufficiently large, the costs can even become negative, implying that $p_b > p^*$ or $p_a < p^*$.

These results derived by STOLL (1978) made use of the assumption that the market maker does know the trade size before quoting the price or that he is allowed to quote different prices for every trade size. In reality, however, the market maker does not know the trade size, he has to accept any trade at a single stated price up to a certain limit.[51] This adds another uncertainty to the market maker, the order size. It can be shown that the results above do not change if the order size is independent of the costs and expected return of the asset. The order size becomes a random variable and therefore instead of Q' and Q^2 we have to insert $E[Q']$ and $E[Q^2]$ into the above derived formulas. No further insight can be gained from this generalization.

The costs and reservation prices derived here are those that occur if the market maker has a time horizon of a single trade, i.e. he makes his consideration on a trade-by-trade basis. Such a short-term behavior can be justified if there is a fierce competition between market makers.

---

and the spread is

$$s = p_a - p_b = p^*(c_a + c_b) = p^* z \sigma^2 Q'.$$

In this representation we can see that the higher the fundamental value of the asset is, the higher we expect the spread to be.

[50] If we define the depth of a market in the conventional way as the amount that can be traded at a given price, we see that the depth and the spread exhibit a positive relationship. The higher the depth of a market, i.e. the larger Q' is, the larger the spread and vice versa.

[51] On the NASDAQ Rule 4613 requires a market maker to accept orders of at least an equivalent of USD 50,000-100,000, depending on several characteristics of the asset and the market, see also chapter 3.5.

If the market maker has a time horizon longer than a single trade, the costs are reduced. By accepting a trade with which the inventory position becomes less favorable there exists the chance that in one of the next trades an offsetting order arrives, what reduces his costs. Nevertheless the influences of the above parameters on the costs do not change significantly.[52]

After having derived the costs and reservation prices of a market maker, we will in the following sections discuss the price setting of market makers, first under competition and then for a monopolistic market maker.

4.3.2 Competitive price setting

Let us for simplicity assume throughout this section that all market makers have the same risk aversion and the order size is fixed to Q', i.e. the spread quoted by all market makers is identical. If all market makers have the same inventory, we can easily see from (4.159) - (4.164) that they all have the same costs and reservation prices. Hence, if they act competitively, they all will quote their reservation prices. Quoting a lower bid or a higher ask price would make this market maker not to offer the best price and hence he is excluded from the order flow by the assumption of strict price priority. He would not receive a trade and thereby make no profits, like in the case when he quotes his reservation price. Quoting a higher bid or a lower ask price would cause a loss, so that he will not quote such prices.

After a single trade however, the inventory position of the market maker executing the trade will change and therewith the costs and reservation prices. Ho and Stoll (1980) provide a framework how competitive market makers set their quotes in such a situation.

Figure 4.13 illustrates the situation where two market makers, A and B, are in the market and have different costs. Suppose the price setting of the bid price, the market maker with the highest bid price receives an incoming order, if both quote

[52] See Stoll (1978, pp. 1148 ff.).

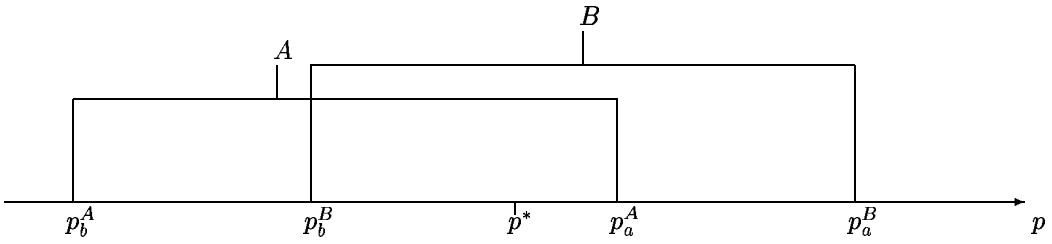


Figure 4.13: Competitive price setting

the same price it is assigned to one of them randomly. Market maker B has the lower costs for a trade at the bid, but by quoting his reservation price he would make no profits. If he lowers his quote he would be able to make a profit. As the costs of market maker A do not allow him to quote a higher price than p_b^A he cannot prevent him from doing so, market maker B still has the best price in the market and will receive all orders until he quotes a price just a fraction above p_b^A. As quoting a price just a fraction above p_b^A gives him the highest profits, he will quote this price.[53] By quoting the same price, p_b^A, he would have to share the order flow with the other market maker, reducing his expected profits.[54] For simplicity we neglect the fraction that has to be quoted above p_b^A and say that he will quote p_b^A, only keeping in mind this arbitrary small fraction.

Obviously, with the same risk aversion it is not possible for a market maker to quote the best price as well on the ask as the bid side, unless all market makers have the same costs and quote their reservation prices. An inventory position allowing him to have smaller costs on one side of the trade gives rise to larger costs on the other side as can easily be seen from (4.159) and (4.160). Generalizing the example from two to M market makers, we derive easily in a similar way the result that the market maker with the lowest costs for this side of the trade receives the order. He quotes

[53] We assume here that prices can be set continuously for simplicity. In case of discrete prices he would have to quote the last discrete price above p_b^A but below p_b^B. We also rule out the possibility that the profit maximum lies between p_b^A and p_b^B as a result of λ_b falling with p_b. In this case we get a behavior comparable to a monopolistic market maker to be treated in chapter 4.3.3.

[54] The possibility that both market makers collude to achieve higher profits by quoting prices that are below the reservation prices of both market makers, is investigated extensively in chapter 5.

the reservation price (minus a fraction) of the market maker with the second lowest costs. If we define C^1 and C^2 as the costs of the market maker with the lowest and second lowest costs, respectively, the profits from a trade are given by

$$
\begin{aligned}
(4.167) \qquad \pi_b^1 &= C_b^1 - C_b^2 \\
&= z\sigma^2 \left(\frac{1}{2}Q^2 + Q'I^1\right) - z\sigma^2 \left(\frac{1}{2}Q^2 + Q'I^2\right) \\
&= z\sigma^2 Q' \left(I^1 - I^2\right), \\
\pi_a^1 &= C_a^1 - C_a^2 \\
&= z\sigma^2 \left(\frac{1}{2}Q^2 - Q'I^1\right) - z\sigma^2 \left(\frac{1}{2}Q^2 - Q'I^2\right) \\
&= z\sigma^2 Q' \left(I^2 - I^1\right).
\end{aligned}
$$

therewith the spread observed in the market (*market spread*) is the difference between the reservation prices of the market makers with the second lowest costs on each side of the trade. Following Ho and Stoll (1983) we will investigate how the market spread is related to the reservation spread. Until now we assumed the market makers to be passive, i.e. to wait for an order arriving at the market, to offset any undesirable inventory position. Another possibility for the market makers would be to initiate a trade by themselves with another market maker (*interdealer trading*).

The expected utility to wait for an offsetting order to arrive on the market is for the market maker with the lowest costs (the other market makers will not receive a trade)

$$
(4.168) \qquad E[U(W_1)] = E[U(W_0)] + U'(W_0)\lambda_a(C_a^2 - C_a^1),
$$

where W_1 denotes the wealth of the market maker in the next period, W_0 the initial wealth, λ_a the probability that an order arrives on the market in the next period and $C_a^1 - C_a^2$ the gain from the trade according to (4.167). This gain in wealth is transformed into utility by multiplying with the first derivative of the utility function using the concept of a first order Taylor series approximation.[55]

[55] The same argumentation can be applied for an offsetting order arriving at the bid.

For interdealer trading we assume that it takes place at the beginning of a period and that the market maker has enough time to update his quotes for his new inventory and then waits for orders to arrive at the market. We further assume that the market maker with the second lowest costs quotes his reservation price.

The expected utility of wealth after interdealer trading, W_0', is composed of the expected utility from his initial wealth, W_0, and an adjustment for the fee he has to pay the other market maker, C_b^2, and the gain from offsetting his inventory, $-C_a^1$:

$$E[U(W_0')] = E[U(W_0)] + U'(W_0)(-C_a^1 - C_b^2). \tag{4.169}$$

After this interdealer trade his inventory has changed from I^1 to $I^1 - Q$ and therewith costs have changed from C_a^1 to $C_a^{1'}$. The expected utility from a trade is now

$$E[U(W_1')] = E[U(W_0')] + U'(W_0)\lambda_a(C_a^2 - C_a^{1'}). \tag{4.170}$$

If we assume W_0 and W_0' to differ not too much such that $U'(W_0) = U'(W_0')$, we get from (4.169) and (4.170):

$$E[U(W_1')] = E[U(W_0)] + U'(W_0)\left((C_a^2 - C_a^{1'})\lambda_a - C_a^1 - C_b^2\right). \tag{4.171}$$

For choosing interdealer trading rather than waiting for an offsetting order it is necessary that $E[U(W_1')] > E[U(W_1)]$. Inserting from (4.168) and (4.171) we get after rearranging:

$$\lambda_a(C_a^2 - C_a^{1'}) < \lambda_a(C_a^2 - C_a^1) - C_a^1 - C_b^2.$$

Inserting for C_a^1 and $C_a^{1'}$ this becomes

$$C_b^2 < z\sigma^2\left(I_1Q' - Q^2\left(\lambda_a + \frac{1}{2}\right)\right). \tag{4.172}$$

If there are only two market makers the market maker who wants to trade with his only competitor will be charged the fee that equals the second best reservation price, i.e. his own reservation price at the bid. Inserting from (4.159) for C_b^2 we get

$$z\sigma^2\left(I_1Q' + \frac{1}{2}Q^2\right) < z\sigma^2\left(I_1Q' - Q^2\left(\lambda_a + \frac{1}{2}\right)\right)$$

which solves for

$$\lambda_a > 1.$$

As λ_a is the probability that an order arrives at the market, it cannot exceed 1, hence with two market makers no interdealer trading occurs.

Having more than two market makers we get with $C_b^2 = z\sigma^2\left(I_2Q' + \frac{1}{2}Q^2\right)$ from (4.172):

$$z\sigma^2\left(I_2Q' + \frac{1}{2}Q^2\right) < z\sigma^2\left(I_1Q' - Q^2\left(\lambda_a + \frac{1}{2}\right)\right),$$

which solves for

$$I_1 - I_2 > Q'(1 + \lambda_a), \tag{4.173}$$

i.e. if the difference in the inventory between the best and the second best market maker is large enough, interdealer trading is induced. For the second best market maker we find $\lambda_a = 0$ as he will never serve a trade, therewith he induces a trade only if $I_1 - I_2 > Q'$. For the best market maker we find $0 \leq \lambda_a \leq 1$, hence for him the divergence has to be even larger before he initiates a trade as he can hope to offset his inventory by an incoming order. If $I_1 - I_2 > 2Q'$ he will always induce an interdealer trade.

If there is the situation that for all market makers $I_1 - I_2 < Q'(1 + \lambda_a)$, i.e. no market maker wants to induce an interdealer trade, there will be no interdealer trading in the future provided that the parameters, Q' and λ_a do not change. To see this, suppose that all market makers have the same inventory, an order arriving will lead to a deviation in the inventory of Q' from the other inventories, no market maker wants to induce a trade. Further orders arriving on the market affect only the inventories of the market makers with the most deviating inventories as they quote the best prices. Furthermore, they will only receive orders offsetting his inventory position, as on the other side he faces too high costs and therefore will not receive an order. Future trades therewith can only narrow the divergence in inventories and no interdealer trading takes place.

The market spread, s_M, is the difference between the best quoted prices on each side of the trade. The best quoted prices are the reservation prices of the second best market makers, hence we have

$$\begin{aligned} s_M &= C_a^2 + C_b^2 \\ &= z\sigma^2\left(\frac{1}{2}Q^2 - I_1Q'\right) + z\sigma^2\left(I_2Q' + \frac{1}{2}Q^2\right) \\ &= z\sigma^2Q^2 + z\sigma^2Q'(I_2 - I_1) \\ &= s + z\sigma^2Q'(I_2 - I_1). \end{aligned} \tag{4.174}$$

With only two market makers it has been shown that no interdealer trading occurs, hence $0 < I_2 - I_1 \leq Q'$, and we find that

$$s \leq s_M \leq 2s. \tag{4.175}$$

With three market makers, one of them is the best at the ask, one at the bid and the third is the second best on both sides, hence $I_2 = I_1$ and we get

$$s_M = s. \tag{4.176}$$

With more than three market makers, the second best market maker on the bid side will have a lower inventory than the second best at the ask side, i.e. $I_2 \leq I_1$. As the difference is allowed to be maximal Q' we find that with more than three market makers

$$0 \leq s_M \leq s. \tag{4.177}$$

Incoming orders have the tendency to balance the inventory of the market makers as always the market maker with the most deviating inventory serves the order. Therewith we have the tendency of $I_2 - I_1$ to converge to zero and the market spread to converge to the reservation spread. If the inventories are quite similar another order increases the divergence and the process of convergence can start again.

Allowing the trade size to vary, adds many more possibilities how the spread can behave over time, but the general finding of the mechanism is not changed.

As all market makers have to pose their quotes for a period simultaneously, it has implicitly been assumed that market makers know the inventory positions of their competitors and can therewith calculate their reservation prices and set prices accordingly, especially the reservation price of the second best market maker has to be known. By observing past prices of the other market makers it is also possible to determine the inventory position by inverting equation (4.159) and (4.160) without initially knowing their inventory position.

If other factors also affect the quoted prices and reservation prices, the exact inventory position cannot easily be determined. It may be impossible to infer the exact reservation prices. If the market makers can infer only the probability distribution of the reservation prices instead of the exact reservation prices, BIAIS (1993) has shown that the bid and ask quotes, on average, are identical to those quoted with full knowledge of the other market makers' reservation prices. He further finds that in this case the spread and the quoted prices are more volatile than with full knowledge of the reservation prices.

After having investigated the competitive price setting, we now turn to the price setting of a monopolistic market maker.

4.3.3 The price setting of a monopolistic market maker

In several markets, e.g. the NYSE, a single market maker is granted a monopoly in providing his services. The main focus of a monopolistic market maker is not to cover his costs, but to maximize his expected utility of terminal wealth by choosing optimal bid and ask prices. HO AND STOLL (1981) provide a model, how a monopolistic market maker sets his prices.

As the demand for the service of the market maker will have an important role, at first we will model this side in more detail. The order arrival rates λ_a and λ_b can be interpreted as the probability that an order arrives on the market within a given

time period $[t, t+1[$. Approximating by a first order Taylor series around zero gives us

$$\lambda_a(p_a) = \lambda_a(0) + \frac{\partial \lambda_a(0)}{\partial p_a} p_a, \tag{4.178}$$

$$\lambda_b(p_b) = \lambda_b(0) + \frac{\partial \lambda_b(0)}{\partial p_b} p_b. \tag{4.179}$$

If we denote the quoted prices to be the fundamental value of the asset, adjusted by the fees the market maker charges, x_a and x_b, respectively, we get in analogy to (4.163) and (4.164):

$$p_a = p^* + x_a, \tag{4.180}$$

$$p_b = p^* - x_b. \tag{4.181}$$

Inserting these relations into (4.178) and (4.179) we can write these relations as

$$\lambda_a(p_a) = \lambda_a(0) + \frac{\partial \lambda_a(0)}{\partial p_a} p^* + \frac{\partial \lambda_a(0)}{\partial p_a} x_a, \tag{4.182}$$

$$\lambda_b(p_b) = \lambda_b(0) + \frac{\partial \lambda_b(0)}{\partial p_b} p^* - \frac{\partial \lambda_b(0)}{\partial p_b} x_b. \tag{4.183}$$

The first terms can be interpreted as the demand of the market maker if he would charge no costs, i.e. it is a measure for the size of the liquidity event investors face. We will denote these terms by α_a and α_b, respectively. The last term can be interpreted as an adjustment in the demand due to the sensitivity of the demand to fees charged, the absolute values of this sensitivities will be denoted β_a and β_b, respectively:

$$\lambda_a(p_a) = \alpha_a - \beta_a x_a, \tag{4.184}$$

$$\lambda_b(p_b) = \alpha_b - \beta_b x_b. \tag{4.185}$$

We assume that there are T time periods in which trading can take place, at every point of time $t \in \{0, 1, \ldots, T\}$ the market maker chooses bid and ask fees that are optimal in the sense that they maximize his expected utility of terminal wealth. At time T the asset is liquidated at the fundamental value and the proceedings are consumed. As before the total wealth of the market maker consists of two

components, his inventory and money:[56]

$$W_t = I_t + M_t \tag{4.186}$$

for all $t = 1, \ldots, T$. The inventory changes with the rate of return and with trading. By assuming again a fixed trade size Q' we get

$$\Delta I_{t+1} = R_{t+1} I_t \Delta t + p^* Q_b + p^* Q_a, \tag{4.187}$$

where Q_b equals Q' if a trade at the bid occurs and is zero otherwise. Q_a equals Q' if a trade at the ask occurs and is zero otherwise. Money holdings change with

$$\Delta M_{t+1} = p_b Q_b + p_a Q_a. \tag{4.188}$$

Let W denote the terminal wealth of the market maker, he then has to maximize $E[U(W)]$ by choosing optimal fees x_a and x_b. We can define a performance function J as

$$J(t, M_t, I_t) = \max_{x_a, x_b} E[U(W)|t, M_t, I_t]. \tag{4.189}$$

After the last trade at time $t = T$ has taken place, the portfolio is liquidated. Therefore the market maker faces no uncertainty and (4.189) has to fulfill the boundary restriction that

$$J(T, M_T, I_T) = U(W_T). \tag{4.190}$$

With optimal fees no further increase in J can be achieved by changing fees, i.e.

$$\max_{x_a, x_b} dJ(t, M_t, I_t) = 0. \tag{4.191}$$

From the principle of optimality in dynamic programming we can use the fundamental recurrence relation[57] and rewrite (4.189) as

$$\begin{aligned} J(t, M_t, I_t) \;=\; & \max_{x_a, x_b} \{ L(t, M_t, I_t, x_a, x_b)\Delta t \\ & + J(t + \Delta t, M_t + \Delta M_{t+1}, I_t + \Delta I_{t+1}) \}, \end{aligned} \tag{4.192}$$

[56] Ho and Stoll (1981) also consider the optimal portfolio as part of the wealth. But as he is compensated for the risks associated with holding the optimal portfolio by the market, it is not necessary to take into account this part of his wealth, but only deviations.

[57] Appendix D.5 gives a brief introduction into dynamic programming.

where $L(t, M_t, I_t, x_a, x_b)$ denotes the expected gain in utility in the current period and J the performance in the remaining periods. The expected gain in utility in the present period consists of the expected gain from a trading at the bid and at the ask, what can be interpreted as the difference in the performance functions with and without a trade:

$$\begin{aligned} L(t, M_t, I_t, x_a, x_b) &= \lambda_a \left(J(t, M_t + p_a Q', I_t - Q') - J(t, M_t, I_t)\right) \\ &\quad + \lambda_b \left(J(t, M_t - p_b Q', I_t + Q') - J(t, M_t, I_t)\right). \end{aligned} \tag{4.193}$$

We can further approximate $J(t+\Delta t, M_t+\Delta M_{t+1}, I_t+\Delta I_{t+1})$ by a first order Taylor series around (t, M_t, I_t):

$$\begin{aligned} J(t + \Delta t, M_t \Delta M_{t+1}, I_t + \Delta I_{t+1}) &= J(t, M_t, I_t) + J_t \Delta t \\ &\quad + J_M \Delta M_{t+1} + J_I \Delta I_{t+1}, \end{aligned} \tag{4.194}$$

where the subscripts denote the partial derivatives with respect to this variable evaluated at the appropriate point. Inserting (4.193) and (4.194) into (4.192) we get after dividing by Δt and rearranging:

$$\begin{aligned} -J_t &= \max_{x_a, x_b} \Bigg\{ J_I \frac{\Delta I_{t+1}}{\Delta t} + J_M \frac{\Delta M_{t+1}}{\Delta t} \\ &\quad + \lambda_a \left(J(t, M_t + p_a Q', I_t - Q') - J(t, M_t, I_t)\right) \\ &\quad + \lambda_b \left(J(t, M_t - p_b Q', I_t + Q') - J(t, M_t, I_t)\right) \Bigg\}. \end{aligned} \tag{4.195}$$

If we let $\Delta t \to 0$, we see that

$$\frac{\Delta I_{t+1}}{\Delta t} \to \frac{\partial I_t}{\partial t}, \tag{4.196}$$

$$\frac{\Delta M_{t+1}}{\Delta t} \to \frac{\partial M_t}{\partial t} = 0. \tag{4.197}$$

As with Δt also λ_a and λ_b converge to zero, the probability of a trade arriving within a very short period of time also approaches zero. We now assume that the return of the risky asset in a given period of time follows a random walk with drift:

$$R_{t+1} = \mu dt + \sigma dz \tag{4.198}$$

with dz denoting a standard Wiener process. Therefore we have from (4.187) with $\Delta t \to 0$ that the change in inventory follows an Itô process:[58]

$$dI_t = \mu I dt + \sigma I dz. \tag{4.199}$$

From Itô's lemma we obtain

$$dJ(t, M_t, I_t) = J_I dI_t + J_t dt + \frac{1}{2} J_{II} (dI_t)^2, \tag{4.200}$$

where

$$(dI_t)^2 = \sigma^2 I_t^2 dt, \tag{4.201}$$

$$J_I dI_t = \mu I_t J_I dt. \tag{4.202}$$

As time is no own variable in the model we find

$$J_t dt = 0. \tag{4.203}$$

Inserting (4.201) - (4.203) into (4.200) gives after dividing by dt:

$$J_t = \mu I_t J_I + \frac{1}{2} J_{II} \sigma^2 I_t^2. \tag{4.204}$$

Using these results we can rewrite (4.195) as

$$\begin{aligned} -J_t = {} & \mu I_t J_I + \frac{1}{2} J_{II} \sigma^2 I_t^2 \\ & + \max_{x_a, x_b} \{ \lambda_a \left(J(t, M_t + (p^* + x_a) Q', I_t - Q') - J(t, M_t, I_t) \right) \\ & + \lambda_b \left(J(t, M_t - (p^* - x_b) Q', I_t + Q') - J(t, M_t, I_t) \right) \} . \end{aligned} \tag{4.205}$$

We denote τ as the remaining time or time horizon, i.e. $\tau = T - t$. As a result we have

$$J_t = J_\tau \frac{\partial \tau}{\partial t} = -J_\tau. \tag{4.206}$$

Define further $LJ = \mu I_t + \frac{1}{2} \sigma^2 I_t^2 J_{II}$, $BJ = J(t, M_t + p^* Q', I_t - Q')$, $SJ = J(t, M_t - p^* Q', I_t + Q')$ and $J = J(t, M_t, I_t)$. Using this notation we can approximate $J(t, M_t +$

[58] See ITÔ (1951) for the definition of such processes and its properties. INGERSOLL (1987, ch. 16) also provides an introduction of Itô processes and shows how to derive Itô's lemma.

$(p^* + x_a)Q', I_t - Q')$ and $J(t, M_t - (p^* + x_b)Q', I_t + Q')$ by a first order Taylor series around $(t, M_t + p^*Q', I_t - Q')$ and $(t, M_t - p^*Q', I_t + Q')$, respectively:

$$
\begin{aligned}
(4.207) \quad J(t, M_t + p_a Q', I_t - Q') &= J(t, M_t + p^* Q', I_t - Q') + J_M x_a Q' \\
&= BJ + J_M x_a Q', \\
(4.208) \quad J(t, M_t - p_b Q', I_t + Q') &= J(t, M_t - p^* Q', I_t + Q') + J_M x_b Q' \\
&= SJ + J_M x_a Q'.
\end{aligned}
$$

We can write (4.205) as

$$
\begin{aligned}
(4.209) \qquad J_\tau &= LJ + \max_{x_a, x_b} \{\lambda_a (BJ - J) + \lambda_b (SJ - J) \\
&\quad + \lambda_a x_a Q' B J_M + \lambda_b x_b Q' S J_M\}.
\end{aligned}
$$

Inserting for λ_a and λ_b we can conduct this maximization and obtain the following first order conditions for the optimal fees x_a and x_b:

$$
\begin{aligned}
(4.210) \qquad \alpha_a - 2\beta_a x_a Q' B J_M - \beta_a (BJ - J) &= 0, \\
\alpha_b - 2\beta_b x_b Q' S J_M - \beta_b (SJ - J) &= 0.
\end{aligned}
$$

The second order condition for a maximum can be shown to be fulfilled as the Hesse-Matrix is negative definite. Solving for the optimal fees we get:

$$
\begin{aligned}
(4.211) \qquad x_a &= \frac{\alpha_a}{2\beta_a} + \frac{J - BJ}{2Q' B J_M}, \\
x_b &= \frac{\alpha_b}{2\beta_b} + \frac{J - SJ}{2Q' S J_M}.
\end{aligned}
$$

The first term can easily be verified to be the optimal fee at which benefits are maximized, provided that a predetermined trade occurs with certainty. The second term adjusts the fees for the risk in the return of the asset and the uncertainty of the order flow. To derive an explicit expression for the fees we have to find a solution for J. Ho and Stoll (1981) provide a separate mathematical appendix where they show how to derive a solution for J by further Taylor series approximations. We do not track down this long and tedious derivation, but only state the results they achieve if the time horizon is not too long.

The fees a monopolistic market maker charges to compensate for the risks he faces, i.e. the second terms in (4.211), depend positively on his risk aversion, the variance of the fundamental value and the trade size. The expressions are similar to those obtained when deriving the costs of market making. As a fourth factor the profits he can make, influence his fees, the larger the expected profits are, the higher fees he charges. These profits depend on the size of the liquidity event (α_a and α_b) and the sensitivity to a higher fee (β_a and β_b). A final factor is the time horizon. The longer the time horizon the higher the fees. With a longer time horizon the market maker faces a higher total risk from holding the inventory and is exposed to the risk of receiving orders that further increase his inventory, this effect more than compensates the benefits from the possibility of an offsetting order arriving at the market.

By comparing the results of the price setting for a monopolistic market maker and competitive market makers, it can be seen that the main difference lies in the fact that a monopolistic market maker is not only compensated for the risk directly connected to his inventory, but also for the risks of the future order flow, i.e. for the risk of an unfavorable shift in his inventory. As this risk increases the longer his time horizon is, the higher the fee he charges. It can be shown that the fees charged by a monopolistic market maker always exceed those of a competitive market maker. Furthermore, the fee is increasing in the level of demand, α_a and α_b, respectively.

4.3.4 Empirical implications

When investigating asset returns on a trade-by-trade basis, it turns out that the first order autocorrelation of the returns are very significantly negative.[59] This effect can easily be explained with the model presented above.

Suppose a very restrictive version of determining the inventory costs of market making by assuming that the inventory of every market maker is zero and that

[59] See e.g. COHEN ET AL. (1980).

therewith all market makers quote their reservation prices.[60] From (4.159), (4.160) and (4.165) we easily see that in this case

$$C_a = C_b = \frac{1}{2}s. \tag{4.212}$$

We further assume that the probability of a trade at the ask and at the bid is both .5. This assumption is reasonable if the fees are equal on both sides of the trade and there is no asymmetric liquidity event favoring trades at one side. Define

$$o_t \sim \begin{cases} 1 & \text{with probability } .5 \\ -1 & \text{with probability } .5 \end{cases}. \tag{4.213}$$

therewith the transaction price is

$$p_t = p^* + o_t\frac{s}{2}. \tag{4.214}$$

Assuming that p^* does not vary over time or we know p^* at every point of time and can eliminate the effect a change in p^* causes and that the inventory does not change over time,[61] we get the price change as

$$\Delta p_t = p_t - p_{t-1} = (o_t - o_{t-1})\frac{s}{2}. \tag{4.215}$$

As $E[o_t] = 0$, we have $E[\Delta p_t] = 0$ and obtain

$$\begin{aligned} Var[\Delta p_t] &= E[\Delta p_t^2] \\ &= \frac{s^2}{4}E[(o_t - o_{t-1})^2] \\ &= \frac{s^2}{4}(0^2\frac{1}{4} + 2^2\frac{1}{4} + 2^2\frac{1}{4} + 0^2\frac{1}{4}) \\ &= \frac{s^2}{2}, \end{aligned} \tag{4.216}$$

$$\begin{aligned} Cov[\Delta p_{t-1}, \Delta p_t] &= E[\Delta p_{t-1}\Delta p_t] \\ &= \frac{s^2}{4}E[(o_{t-2} - o_{t-1})(o_{t-1} - o_t)] \\ &= \frac{s^2}{4}(0 - 0 - (1\frac{1}{4} + 1\frac{1}{4} + 1\frac{1}{4} + 1\frac{1}{4}) + 0) \\ &= -\frac{s^2}{4}. \end{aligned} \tag{4.217}$$

[60] This method has first been presented in ROLL (1984).

[61] This assumptions seems to be very restrictive, because we know from above that the inventory changes after every trade. CAMPBELL ET AL. (1997, p. 102) report that by allowing inventories to vary over time the result is not much affected as the spread quoted by a single market maker does not vary with his inventory. A simulation undertaken by KRAUSE (1999) verifies this results in a much more general framework.

therewith the first order autocorrelation is

$$\rho_1 = -\frac{1}{2}. \tag{4.218}$$

Higher order autocorrelations can easily be shown to be zero. By the possibility of trading one period at the ask and the other period at the bid the observed trade data are negatively correlated. This result can help to understand the empirical findings of negatively autocorrelated price changes or returns.[62]

Until a few years ago it was not possible to receive data on each trade and the associated quotes of market makers. With only daily data available, e.g. closing prices, the spread has been estimated by inverting (4.217):

$$s = 2\sqrt{-Cov[\Delta p_{t-1}, \Delta p_t]}. \tag{4.219}$$

As the covariance is also influenced by other factors, e.g. the information flow or competition between market makers, it has been not a very liable technique to estimate the spread. Various modifications have been developed to address these problems. Since nowadays data on each date are available, there is no longer a need for applying these methods.

4.4 Information-based models of market making

The last section focused on the influence of inventory costs on bid and ask prices and the spread. It has been assumed that all investors and market makers agree on the fundamental value of the asset, i.e. have equal information. In this section we will take up the line already laid in section 4.2 by assuming that there exist two groups of investors, informed and uninformed investors. Unlike in the models building on KYLE (1985), we will not aggregate the order flow over a given period of time, but examine the behavior on a trade-by-trade basis by introducing a market maker replacing the match maker.

[62] See CAMPBELL ET AL. (1997, pp. 101 f.). A similar result can also be obtained for a monopolistic market maker.

The basis for information-based models of market making has been laid by BAGEHOT (1971) with his distinction of market and trading gains. A *market gain* arises from the price change and dividends of an asset in a given time period. An investor can realize the market gain by holding the asset during the entire time period without trading. Let p_1 denote the price of the asset at the end of the period and p_0 at the beginning, then with assuming that no dividends are paid, the market gain is

$$\Delta p = p_1 - p_0. \tag{4.220}$$

The market gain is the same for all investors holding the asset at the beginning and at the end of the period. Let us now assume that all investors hold the same amount of assets at the beginning and at the end of the period.[63] Investors can either only hold the asset over the entire period or they can trade with each other. The gain associated with this trading activity is the *trading gain*, denoted f^i. The total gain of investor i is

$$\pi^i = \Delta p + f^i. \tag{4.221}$$

The total gain of all investors can only be the market gain by comparing the wealth at the beginning and at the end of the period, i.e. with N investors it is

$$\pi = \sum_{i=1}^{N} \pi^i = \sum_{i=1}^{N} \Delta p = N\Delta p. \tag{4.222}$$

By aggregating (4.221) over all investors we get

$$\pi = \sum_{i=1}^{N} \pi^i = \sum_{i=1}^{N} (\Delta p + f^i) = N\Delta p + \sum_{i=1}^{N} f^i. \tag{4.223}$$

Inserting from (4.222) we receive

$$\sum_{i=1}^{N} f^i = 0, \tag{4.224}$$

i.e. the total trading gains are zero. For every investor making trading profits, there must be another making a loss. If there are two groups of investors, informed

[63] If some investors decide to liquidate their position in the asset, there has to be found another investor taking his position as the number of outstanding shares is fixed. We therefore could aggregate these investors such that the assumption is fulfilled.

and uninformed, where uninformed investors have to trade for exogenous reasons, informed investors will only trade if they expect to profit from trading, while the uninformed are forced to trade even if they make trading losses.

With a market maker acting as counterpart in every trade, the informed investors will only trade with him if he sets a price at which they make an expected profit, otherwise they would refuse to trade. Hence, if we allow only for a single trade per period of time, we see from (4.224) that with only two market participants, the market maker will make a loss from trading with an informed investor. He will try to reduce the loss he receives from trading with an informed investor by quoting a less favorable price than what he thinks the fundamental value is, but he will never be able to make a gain. This remaining loss he has to offset from another source. This source are uninformed investors. He has to charge a price to them such that he makes a gain from trading with them and they make a loss.[64]

By the anonymity of the two investors provided by a broker, the market maker does not know whether the investor he trades with is informed or uninformed. We will see below that this problem of loosing to one and gaining from the other group, gives rise to the spread. The losses of uniformed to informed investors are also called *adverse selection costs*.[65]

BAGEHOT (1971) further states that market makers will be uninformed. They observe the order flow and try to balance the buy and sell orders by quoting appropriate prices. By observing the order flow they aggregate the information available in the market as well as the errors. Errors cancel out by the law of large numbers if the number of informed investors is sufficiently high and they make no systematic errors. If the market maker would become informed and quote his prices according to his own information, he faces the risk of relying on a large error in his consider-

[64] If informed investors would buy the asset, he can hope that an uninformed investors arrives on the market and sells the asset to him. As he charges a less favorable price to him than his inference of the fundamental value, he makes a profit from this trade.

[65] The problem of asymmetric information in markets was first formalized in AKERLOF (1970). He pointed out that asymmetric information between market participants trading with each other, imposes costs on the less well informed and may even lead to a breakdown of the market.

ations, causing him a large loss. Therefore he will not invest his wealth to become informed.[66]

4.4.1 Determination of adverse selection costs

The first to formalize the idea of BAGEHOT (1971), were COPELAND AND GALAI (1983). They provide a simple framework in which only a single trade takes place before the information is fully revealed to all market participants, e.g. by liquidation of the asset at the fundamental value. The liquidation value of the asset is a random variable, p, which has a known distribution F and an expected value of $E[p] = p_0$. It is assumed that informed investors know the exact liquidation value, p^*, before trading takes place. Uninformed investors trade for exogenous reasons, but their demands depend on the fee charged by the market maker, the larger the fee, the smaller their demand, hence they are hedgers. The market maker is assumed to be risk neutral, i.e. he faces no inventory costs.

Trading takes place as follows: At the beginning informed investors become to know the liquidation value and the market maker sets his prices. Knowing these prices informed and uninformed investors place their orders. Only one of these orders will be served by the market maker, this order is chosen randomly. The probability that the order chosen is from an informed investor is γ_I. This probability depends on the prices he quotes. The larger the fee from the point of an uninformed investor, the less uninformed investors submit orders and hence γ_I increases. An informed investor will submit a buy order if $p_a < p^*$ and a sell order if $p^* < p_b$. If $p_b < p^* < p_a$ he will not submit an order and all orders in the market are from uninformed investors. As the market maker does not know p^*, he does not know what his loss is, he only knows that he makes a loss. His expected profit from trading with an informed

[66] Although the assumption of uninformed market makers is common in dealer markets, it can reasonably be assumed that market makers can get private information from other sources, such that they will not be completely uninformed. CALCAGNO AND LOVO (1998) provide a model where market makers are at least partially informed and have different information. However, we consider this line of research not further at this point as it gives no substantial new insights into the behavior of market makers.

investor is

$$E[\pi_I] = \int_{p_a}^{\infty} (p_a - p) dF(p) + \int_0^{p_b} (p - p_b) dF(p) \leq 0. \tag{4.225}$$

If he trades with an uninformed investor the agreed value of the asset is p_0, he will make a profit of $p_a - p_0$ from a trade at the ask and of $p_0 - p_b$ from a trade at the bid. Uninformed investors can either buy or sell the asset, we assign a probability of γ_a for a trade at the ask and γ_b for a trade at the bid, such that $\gamma_a + \gamma_b = 1$. These probabilities depend on the fees charged by the market maker, or equivalently by the prices he sets, such that with $\gamma_U = 1 - \gamma_I$ as the share of uninformed investors in the market, we can reasonably assume

$$\begin{aligned} \frac{\partial \gamma_U}{\partial p_a} &> 0, & \frac{\partial \gamma_U}{\partial p_b} &< 0, \\ \frac{\partial \gamma_a}{\partial p_a} &< 0, & \frac{\partial \gamma_b}{\partial p_b} &> 0, \\ \frac{\partial \gamma_a}{\partial p_b} &= 0, & \frac{\partial \gamma_b}{\partial p_a} &= 0. \end{aligned} \tag{4.226}$$

The expected profits from trading with an uninformed investor are

$$E[\pi_U] = \gamma_a(p_a - p_0) + (1 - \gamma_a)(p_0 - p_b) \geq 0. \tag{4.227}$$

The total expected profits of the market maker are

$$E[\pi] = \gamma_I E[\pi_I] + (1 - \gamma_I) E[\pi_U]. \tag{4.228}$$

The costs or reservation prices are obtained by setting $E[\pi] = 0$. A monopolistic market maker would maximize (4.228).

Solving the problem to determine the adverse selection costs would rely on many assumptions and is therefore not conducted here. Nevertheless we can use (4.228) to derive some implications of the model. It can be shown that we always have a positive spread if $\gamma_I > 0$, i.e. if there are informed investors. Suppose that

$p_a = p_b = p'$, we then get from (4.225):

$$
\begin{aligned}
(4.229)\quad E[\pi_I] &= \int_{p'}^{\infty}(p'-p)dF(p) + \int_0^{p'}(p-p')dF(p) \\
&= p'\int_{p'}^{\infty} dF(p) - \int_{p'}^{\infty} pdF(p) + \int_0^{p'} pdF(p) - p'\int_0^{p'} dF(p) \\
&= p' - p'\int_0^{p'} dF(p) - E[p] + \int_0^{p'} pdF(p) \\
&\quad + \int_0^{p'} pdF(p) - p'\int_0^{p'} dF(p) - p'\int_0^{p'} dF(p) \\
&= p' - p_0 + 2\int_0^{p'}(p-p')dF(p).
\end{aligned}
$$

As $\int_0^{p'}(p-p')dF(p) < 0$ we get for $p' < p_0$ that $E[\pi_I] < 0$ and as also $\int_0^{p'}(p-p')dF(p) < p_0 - p'$ we have for $p' \geq p_0$ $E[\pi_I] < p_0 - p' < 0$. Hence we always find $E[\pi_I] < 0$. From (4.227) we obtain

$$
(4.230)\qquad E[\pi_U] = \gamma_a(p'-p_0) + (1-\gamma_a)(p_0-p') = (p'-p_0)(2\gamma_a - 1).
$$

If $p' < p_0$ we need $\gamma_a < \frac{1}{2}$ for $E[\pi_U] > 0$, what is necessary to compensate for the loss from trading with informed investors. This would imply that the probability of a trade at the bid is higher than a trade at the ask, although the fee for a trade at the bid is positive while it is negative for a trade at the ask. If we reasonably assume that the probabilities for a trade are equal for the same fee this contradicts (4.226), hence we cannot set $p' < p_0$. With the same argumentation the case $p' > p_0$ can be ruled out. Hence the only solution is to set $p' = p_0$, for which $E[\pi_U] = 0$. Inserting these findings into (4.228) we receive

$$
(4.231)\qquad E[\pi] = \gamma_I E[\pi_I] < 0.
$$

As a market maker will not accept to make a loss, the spread always has to be positive if $\gamma_I > 0$.[67]

Further we can see from (4.228) that the more informed investors are present, i.e. the higher γ_I is, the larger the first term becomes. To compensate the losses from

[67] A negative spread has earlier been pointed out to be not possible by arguments of arbitrage.

a higher probability of trading with informed investors a larger spread has to be quoted.[68]

If the uncertainty about the liquidation value increases, i.e. more probability is put on the tails of F, e.g. through a larger variance, we can see from (4.225) that the expected losses from trading with informed investors increase. This is compensated by the market maker by quoting a larger spread.

The results found are very similar to those of KYLE (1985) in the case of a single auction. As in the Kyle-model it would therefore be interesting to extend this static model to a dynamic model by allowing more than a single trade in a given time period before the asset is liquidated. This generalization has been undertaken by the model of GLOSTEN AND MILGROM (1985).

With only a single trade before the asset is liquidated there is no need to exploit information from the order flow. If there are more trades, information from the order flow will be used by the market maker to minimize his losses from trading with an informed investor through updating his beliefs.

The market maker uses all information available to him from former trades, denoted Ω_t. He knows that an informed investor only trades if the quoted price is above (trade at the bid) or below (trade at the ask) the liquidation value, while an uninformed investor trades independently of the quotes and the fundamental value. We only assume that the frequency with which he trades is sensitive to the fee the market maker charges. His belief on the fundamental value equals that of the market maker as both are uninformed, but the market maker can observe the former order flows.

The probability that a buy or a sell order arrives at the market are

$$\text{(4.232)} \quad \text{Prob}(\text{Buy order}|\Omega_t) = \gamma_I \text{Prob}(p^* > p_a^t|\Omega_t) + (1-\gamma_I)\gamma_a,$$

$$\text{(4.233)} \quad \text{Prob}(\text{Sell order}|\Omega_t) = \gamma_I \text{Prob}(p^* < p_b^t|\Omega_t) + (1-\gamma_I)(1-\gamma_a),$$

[68] In general it cannot be determined how this larger spread is achieved, by increasing only the ask, decreasing only the bid price or a combination of these. ADMATI AND PFLEIDERER (1988) provide an interesting strategy of the market makers in setting their prices as will be presented in section 4.4.4.

where p_a^t and p_b^t denote the ask and bid prices in trading round t, respectively. Let us denote the order form by o_t, where

$$o_t = \begin{cases} 1 & \text{if the order is a buy order} \\ -1 & \text{if the order is a sell order} \end{cases}, \tag{4.234}$$

then $\Omega_t = \{o_1, \ldots, o_{t-1}\}$. The expected profits of the market maker from a trade of a certain size are given by

$$\begin{aligned} E[\pi_t|\Omega_t] &= (p_a^t - E[p^*|\Omega_t, o_t = 1])\text{Prob}(o_t = 1|\Omega_t) \\ &\quad +(E[p^*|\Omega_t, o_t = 1] - p_b^t)\text{Prob}(o_t = -1|\Omega_t). \end{aligned} \tag{4.235}$$

In order to determine the costs of market making his expected profits have to equal zero. He can achieve this either by setting both prices such that he makes no profit on either side or he can set the prices such that he makes a profit on one side of the trade and a loss on the other. If we assume competitive market makers, the profits on one side of the trade will be deteriorated by other market makers undercutting the price by applying another price strategy, hence the second alternative is ruled out as all market makers have the same costs. Therewith the bid and ask prices are determined as follows:

$$\begin{aligned} p_a^t &= E[p^*|\Omega_t, o_t = 1], \\ p_b^t &= E[p^*|\Omega_t, o_t = -1]. \end{aligned} \tag{4.236}$$

As we see from this expression, the bid and ask prices are determined such that they equal the beliefs of the market maker given the new information on the side of the current trade. In the KYLE (1985) model of chapter 4.2.1 we imposed exactly this condition on the price setting behavior of the match maker.

With Bayes rule,[69] (4.232) and (4.233) we get the update of the beliefs as follows, where p^t denotes the transaction price applied in trading round t, i.e. $p^t = p_a^t$ if the trade occurred at the ask and $p^t = p_b^t$ if the trade was at the bid:

$$\begin{aligned} \text{Prob}(p^* > p^t|\{\Omega_t, o_t = 1\}) &= \frac{\text{Prob}(p^* > p_a^t|\Omega_t)\left[\gamma_I + (1-\gamma_I)\gamma_a\right]}{\text{Prob}(p^* > p_a^t|\Omega_t)\gamma_I + (1-\gamma_I)\gamma_a} \\ &\begin{cases} = & \text{Prob}(p^* > p_a^t|\Omega_t) \quad \text{if} \quad \gamma_I = 0 \\ > & \text{Prob}(p^* > p_a^t|\Omega_t) \quad \text{if} \quad \gamma_I > 0 \end{cases}. \end{aligned} \tag{4.237}$$

[69] Appendix C.2 provides a brief introduction to Bayesian learning and Bayes rule.

As long as there are informed investors, a buy order increases the probability that the fundamental value is above the ask. A similar result can be obtained for a trade at the bid, it increases the probability that the fundamental value is below the bid. With this result we get with (4.236)

$$
\begin{aligned}
(4.238) \qquad p_a^t &= E[p_a^t|o_t = 1] = E[E[p^*|\Omega_t, o_t = 1]|o_t = 1] \\
&= E[p^*|\Omega_t, o_t = 1] \\
&\quad \begin{cases} = & E[p^*|\Omega_t] \quad \text{if} \quad \gamma_I = 0 \\ > & E[p^*|\Omega_t] \quad \text{if} \quad \gamma_I > 0 \end{cases},
\end{aligned}
$$

i.e. as long as there is the threat of adverse selection the ask will exceed the fundamental value assigned by the market maker. The same result can be obtained for the bid such that

$$
\begin{aligned}
(4.239) \qquad p_a^t &= E[p^*|\Omega_t] = p_b^t \quad \text{if} \quad \gamma_I = 0, \\
p_a^t &> E[p^*|\Omega_t] > p_b^t \quad \text{if} \quad \gamma_I > 0.
\end{aligned}
$$

If there is the possibility of trading with an informed investor, the spread will always be positive due to adverse selection costs. As we can easily see, (4.237) is increasing in the fraction of informed investors, γ_I. Therewith from (4.238) we see that p_a^t increases and in analogy p_b^t decreases, hence the spread increases in γ_I.

By using (4.236) it can be shown that

$$
(4.240) \qquad E[p^{t+1}|\Omega_t] = E[E[p^*|\Omega_t, o_t]|\Omega_t] = E[p^*|\Omega_t] = E[p^*|\Omega_{t-1}, o_t] = p^t.
$$

This implies that expected price changes between two subsequent trades are independent of each other if the fundamental value does not change. These findings are in contrast to the behavior with inventory costs, where price changes are negatively correlated.

From (4.236), (4.238) and (4.239) we get

$$
\begin{aligned}
(4.241) \qquad p_a^{t+1} &= E[p^*|\Omega_{t+1}, o_{t+1} = 1] \\
&> \begin{cases} E[p^*|\Omega_{t+1}] & \text{if } o_t = 1 \\ E[p^*|\Omega_{t+1}] & \text{if } o_t = -1 \end{cases} \\
&= \begin{cases} p_a^t & \text{if } o_t = 1 \\ p_b^t & \text{if } o_t = -1 \end{cases} \\
p_b^{t+1} &= E[p^*|\Omega_{t+1}, o_{t+1} = -1] \\
&< \begin{cases} E[p^*|\Omega_{t+1}] & \text{if } o_t = 1 \\ E[p^*|\Omega_{t+1}] & \text{if } o_t = -1 \end{cases} \\
&= \begin{cases} p_a^t & \text{if } o_t = 1 \\ p_b^t & \text{if } o_t = -1 \end{cases} .
\end{aligned}
$$

If the previous transaction has taken place at the ask (bid) the new ask (bid) price is higher (lower) than the previous ask (bid) price. We see further that always $p_a^{t+1} > p_b^t$ and $p_b^{t+1} < p_a^t$, i.e. the prices will never be revised so much that both prices are outside the former spread. This price behavior can easily be explained by looking at (4.236). The former ask and bid prices were the best guess of the market maker given the transactions he was waiting for, hence from (4.239) we see that $p_a^{t+1} > E[p^*|\Omega_{t+1}] = p^t > p_b^{t+1}$, what proofs our claim.

Define the spread as $s_t = p_a^t - p_b^t$. Glosten and Milgrom (1985) then show that for the average spread $\overline{s} = \frac{1}{T}\sum_{t=1}^{T} s_t$ we find

$$
(4.242) \qquad E[\overline{s}^2] \leq \frac{\xi Var[p^*]}{T},
$$

where ξ is some constant, T the number of trading rounds and $Var[p]$ the unconditional variance of the fundamental value.

We find the average squared spread to be independent of any trade patterns. We can easily derive a relationship between the average spread and the average trading volume in a given period of time. Assuming that within a given period of time, T trades of a fixed size occur, a large T implies a high trading volume. As can be seen from (4.242) and as ξ and $Var[p^*]$ are independent of T, we expect a small average spread.

With a large number of trades the information of informed investors is revealed much faster to the market maker than with only few trades. This reduces the adverse selection costs for later trading rounds and hence the spread. The initial adverse selection costs at the first trades, represented by $Var[p^*]$ increase the average spread.

As we saw in the HOLDEN AND SUBRAHMANYAM (1992) model presented in chapter 4.2.2, trading will be especially intense short after information has been released to informed investors, in our case the early trades of the time period, causing adverse selection costs to increase with γ_I in the first trades, but then reduces quickly such that they are small on average. The result obtained here can explain why spreads in actively traded stocks are on average smaller than in less actively traded stocks.

KRISHNAN (1992) showed that the models on auction markets described in chapter 4.2 and the model of GLOSTEN AND MILGROM (1985) used here, are equivalent if we restrict the auction models slightly. Although the auction models allowed for no market makers and no spread, the prices at which trades occur turn out to be identical if we allow only for fixed order imbalances of the same size as an order size here. In this case the price determined by the match maker equals the bid and ask prices of the market maker, depending on the sign of the order imbalance. The properties of the prices are identical in both cases, a result confirmed by models from TONKS (1997) and MADHAVAN (1992).

Although the model of GLOSTEN AND MILGROM (1985) model gave further insights into the behavior of bid and ask prices and the spread, no explicit formula could be derived from the model, but the equivalence with the auction models provides us with such formulas to model the dynamics of prices.

4.4.2 Simultaneous trading on different stock exchanges

The result that competitive market makers set prices equal to the expected fundamental value given the order flow is due to the assumption of Bertrand competition

between market makers. Although not explicitly stated in the models presented thus far, it has been assumed that investors route their entire order flow to a single market maker quoting the most favorable price. As market makers quoting less favorable prices do not participate in the order flow, they make zero profits, ignoring fixed costs. Hence Bertrand competition requires market makers to quote prices such that they make expected profits of zero, which has been shown in the previous section to imply quotes that equal market maker's inference about the fundamental value.

In the models considered, the assumption of strict price priority ensured the entire order flow to be routed to a single market maker quoting the best price for the entire order. As has been shown that the quoted prices become less favorable the larger the order size is, it could be profitable for investors to split their orders between market makers to reduce the order size submitted to each market maker and therewith receive more favorable prices.[70] For this reason also market makers quoting not the most favorable price will receive a fraction of the order flow, the less favorable the quotes are, the smaller this fraction will be. The optimal splitting of the order flow will give an investor equal costs for trading with every market maker.

Splitting the order flow requires that investors are able to route their orders directly to different market makers, violating strict price priority. A possibility to achieve such a situation is the assumption that an asset is traded on several stock exchanges and that trading on these stock exchanges can take place simultaneously. A more realistic setting would be to assume a market with multiple market makers. If each of these market makers quotes prices for a fixed maximum order size, e.g. a lot, then larger orders are splitted according to these sizes and executed separately. The first part will be executed by the market maker offering the best price. As the update of quotes takes time, the remaining orders are executed with the market maker offering the second best price, and so forth. This mechanism can be found

[70] According to BERNHARDT AND HUGHSON (1997) it is common to assume that market makers condition their quotes only on the order flow they receive, but not on the order flow the other market makers receive, hence such a splitting of the order flow would not influence the quoted prices.

for example on the NASDAQ. The implicit splitting of orders as described here has the same implications as an explicit order splitting for which the model has initially been designed. given this possibility to split the order flow between market makers, BERNHARDT AND HUGHSON (1997) show that competitive market makers are able to make positive expected profits by quoting less competitive prices.

The intuition behind their result is that even when quoting not the most favorable price, market makers still will receive a fraction of the order flow, whereas if investors are not allowed to split the order flow, they would receive no orders. This situation enables market makers, as formally shown in BERNHARDT AND HUGHSON (1997), to quote less competitive prices, i.e. higher ask and lower bid prices, and make positive expected profits. These less favorable prices increase the trading costs for investors. They further show that prices become more competitive by increasing the number of market makers, in the limiting case, with an infinite number of market makers, the quoted prices are competitive.[71]

Furthermore DENNERT (1993) points out that as a result of splitting the order flow the share of informed trades increases, i.e. they trade more aggressively. Therewith adverse selection costs increase, as shown in the previous section, and hence trading costs for uninformed investors.

With these two effects due to the possibility to split the order flow, competition among market makers not necessarily reduces trading costs as usually expected.

4.4.3 Comparing competitive and monopolistic market makers

If the number of informed investors increases, it has been shown above that the adverse selection costs increase and therewith the fee charged by the market maker. If we assume that uninformed investors' demands respond to the fee charged, their

[71] BERNHARDT AND HUGHSON (1997) also show that with enabling investors to split the order flow, a linear equilibrium as derived in chapter 4.2 only exists if the demand of liquidity traders for trading is price elastic, i.e. uninformed investors are hedgers.

demands reduce further, the share of informed investors increases and the fee charged by market makers increases further. If adverse selection costs are high enough, the demand of uninformed investors nearly vanishes and the market makers cannot offset their losses from trading with informed investors. They have to quote such high fees that the market breaks down.

The market breakdown as the result of too much informed investors is due to the need of the market maker to achieve zero expected profits. It has been shown in chapter 4.2.2 that especially short after new information is available, informed investors trade very actively on this information. Afterwards the information is revealed through prices, the trades of informed investors reduce and with it adverse selection costs. However, if at the beginning the market breaks down, prices never will be able to reveal information and the threat remains *ad infinitum*. A formal model to show this intuition is provided by GLOSTEN (1989). With further assumptions like a normal distributed liquidation value, conditions are derived under which the market breaks down as the result of too high adverse selection costs with competitively acting market makers.

To overcome the problem of a market breakdown, he proposes to establish a monopolistic market maker. As such a market maker maximizes his total profits over a given time period and has not necessarily to avoid losses from every trade, he can be compensated for losses by larger profits in future trades. GLOSTEN (1989) shows that the optimal price setting strategy of a risk neutral monopolistic market maker is to set an average price that enables him to make extraordinary profits in normal trading environments and compensates him for losses in times of high adverse selection costs.

By incurring a loss short after new information has become available to informed investors, he keeps the market open and learns the information through orders he receives. After having learnt this information sufficiently well, his adverse selection costs are reduced and he can make larger profits compensating him for the incurred loss. By shutting the market down, he would not be able to learn the information

and hence would make no profits. In contrast, competing market makers cannot expect to offset their losses by making profits in future trading, hence they are not willing to face losses to learn the information.

As long as adverse selection costs are higher than a certain threshold, a monopolistic market maker quotes prices more favorable to investors than a competitive market maker would be able to do. Only if adverse selection costs are low, competing market makers charge a smaller fee. This result is different to the result obtained for inventory-based models, where a monopolistic market maker always quotes less favorable prices than competitive market makers.

We should therefore find competing market makers in markets with only small adverse selection costs, i.e. only few informed investors, while for markets with high adverse selection costs, i.e. many informed investors, a monopolistic market maker would be preferred.

The NASDAQ has a system of competing market makers, while the NYSE has a monopolistic market maker (specialist). The many analysts following the companies on the NYSE make adverse selection costs much higher on the NYSE than on the NASDAQ, where especially the information of the informed investors are much less precise than on the NYSE, as the companies mostly work in a more dynamic environment. This would be a rationale for the NYSE to have a monopolistic market maker and the NASDAQ to have competing market makers.

4.4.4 Explaining return patterns

There exists significant evidence of systematic patterns in asset returns, as well within a trading day as across trading days. The most prominent of these effects is the Monday-effect, where returns on Mondays are negative on average. Admati and Pfleiderer (1989) present a model that is able to explain such a pattern.

In section 4.2.4 the model of Admati and Pfleiderer (1988) has been presented, that showed the concentration of trading in a few periods. This pattern in trading

volume arises endogenously out of the model as uninformed investors want to reduce their adverse selection costs from trading with informed investors. Such a behavior was able to explain pattern in volatility and trading volume, but not in returns.

By introducing a market maker we are able to explain such patterns in expected returns. As uninformed investors react sensitive to the fee charged by the market maker, quoting a high fee reduces the trading of uninformed investors and increases adverse selection costs. If the market maker quotes a low fee his profits are small, but also his adverse selection costs are reduced by the increased number of uninformed investors trading.

By quoting a very low fee on one side of the trade, e.g. the ask side, and a high fee on the other side of the trade, in our example the bid side, the market maker will induce many uninformed trades on the ask side and virtually a breakdown of the market on the bid side. Although this price setting behavior causes a large order imbalance, the ask price does not change significantly after every trade as most trades are from uninformed investors. On the bid side nearly no uninformed investors trade due to the high costs, nearly all trades come from informed investors, hence an order submitted to the bid side is much more informative than at the ask side. The high fee charged and the information revelation reduces the profits of informed investors from trading on their information. If it is known that the market maker will reverse his strategy in the next period of time, e.g. the next trading day, by quoting low fees at the bid and high fees at the ask side, an informed investor will prefer to wait until this period as his profits are increased, provided that his information is not revealed beforehand by other means. The market maker will have to reverse his strategy to offset his large inventory position he had to acquire as a result of the large order imbalance in the first period.

With this quoting strategy the market maker does not only induce a pattern in trading volume, but also a pattern in the types of trades that occur. In the example above the market maker induces more trades at the ask in the first and more trades at the bid in the second period. Admati and Pfleiderer (1989) show that such a

strategy is an equilibrium as well for a monopolistic market maker as for competitive market makers. They furthermore show that the concentration of trading at the bid and at the ask under certain assumptions does not take place in the same period like in ADMATI AND PFLEIDERER (1988).

If the trading at the ask concentrates in a period and the trading at the bid in the next period, it is more likely to observe an ask price in the first and a bid price in the second period. Therewith the expected return can easily be shown to be negative between these two periods. If in the next period trading again concentrates at the ask, the expected return will be positive.[72]

The model presented here shows that patterns in expected returns can arise, but not how they are timed, neither that they have to be timed equally by all market makers. This timing has to be explained by exogenous coordination of the behavior of market makers. The Monday-effect suggests with this model a concentration of buy orders on Fridays and a concentration of sell orders on Mondays. As over the weekend no trading is possible and there are many investment funds that have to invest money newly acquired, they invest it on Fridays. Over the past years there has been a massive flow into funds, such that they have to invest, i.e. buy. Their need to buy can be justified by the frequent benchmarking with an index, as the closing of an index is taken as their benchmark, they best can track the index by trading near the close, i.e. the end of the day or the end of the week. This causes a concentration of trading at the ask on Fridays, resulting in an expected loss on Mondays as well as for the beginning of the next trading day, an effect that also is confirmed empirically.

[72] This behavior would give rise to returns being negatively correlated. Depending on the fees charged, this correlation will be below or above -.5, which found to be the result of the presence of a spread.

4.5 Liquidity provision with limit orders

In all models presented thus far, investors were only allowed to submit market orders to a match maker or market maker. The market maker has been the sole provider of immediacy. As has already been mentioned in section 2.4, many different order forms, besides market orders, exist, the most important being limit orders.[73] Limit orders can be viewed as an alternative to market makers for providing immediacy. Like the quotes of a market maker it enables investors to trade at a fixed price, the limit price, with certainty. For such an investor there is no difference between a limit order and the quotes of a market maker. We can therefore interpret the quotes of a market maker also as a limit order.

Despite these similarities between market makers and limit orders, the concepts of inventory costs and adverse selection costs, cannot easily be applied to investors submitting limit orders. The most important difference between a market maker and a limit order trader is that the market maker is obliged to quote both bid and ask prices, while the limit order trader is free to submit either a limit order, a market order or not to trade at all. He therefore will only submit a limit order if this is the most profitable alternative for him. Despite these differences their similarity for investors submitting market orders suggests that limit orders are competitors to the quotes of market makers and that they will influence each other. In this section it shall be investigated which consequences arise from the introduction of limit orders.

Throughout this section we assume strict price priority not only between quotes of market makers, but also between their quotes and limit orders and within limit orders, i.e. an incoming market order is executed at the best available price, a limit order or the quote of a market maker.[74]

[73] The importance of limit orders even exceeds that of market orders. According to CHAKRAVARTY AND HOLDEN (1995, pp. 213 ff.), in the second half of 1993 62% of all orders submitted to the SuperDOT system of the NYSE were limit orders, furthermore the market maker has only been involved in 17% of all transactions. HARRIS AND HASBROUCK (1996) report 45% of all orders submitted to the SuperDOT system to be limit orders.

[74] Not all exchanges apply such strict rules. The NASDAQ goes even a step further to enhance competition between limit orders and market makers by forcing market makers to give limit

4.5.1 The placement of limit orders

Despite their importance in trading, limit orders have only recently attracted more attention. One of the few early exemptions are COHEN ET AL. (1981), who investigate the optimal order placement decision of an investor and its implications for the spread.

They do not distinguish between limit orders and quotes of a market maker, as only the decision of an investor is considered. It is assumed that investors are free to submit an order at every time, the order can either be a market or a limit order for a fixed trade size. The arrival of market orders is assumed to follow a Poisson process with order arrival rate λ, i.e. per period of time the expected number of market orders equals λ. A limit order can be submitted at any price, it cancels if the limit order has to execute a market order. The investor is then free to submit a new order, either a limit or market order.[75]

If the limit order is not executed, it remains valid for the whole period of time, e.g. a trading day and cancels automatically afterwards. All limit orders are assumed to be published, i.e. the limit order book is open. Investors see not only the best prices available, but also all other limit orders currently unexecuted.

We assume now an investor who wants to buy the asset.[76] He can do so either by submitting a market buy order that executes with certainty, i.e. probability one, at the best available price p_a. The alternative he has now, is to submit a limit bid order. When submitting a limit bid order, the execution of this order is not guaranteed as it has to wait for a market sell order to arrive on the market. Before such an order executes, there must be no other limit bid orders offering a more favorable price. This uncertainty of execution imposes costs on the investor, which earlier already have been considered as waiting costs, reducing his expected utility from trading by

orders priority at the same price as the quotes of a market maker.

[75] Without changing the argument we can introduce a market maker to ensure at least one limit order on both sides of the market.

[76] The same considerations can be made for an investor selling the asset.

a limit order. But on the other hand he will be able to trade at a more favorable price.

It is obvious from previous sections, the higher he sets his limit bid price p_b^L, the higher the probability of execution, ϕ.[77] If he sets a limit bid price above p_a his order will execute with probability one as investors could make arbitrage profits by selling at p_b^L and buying at the lower price p_a. This negative spread is not rational as the investor would be able to buy the asset at a lower price by submitting a market order. If p_b^L increases from below to p_a the probability of execution increases, but does not converge to one. This can easily be proved as follows.

The number of trades in a time period in which the limit order remains valid, $N(\lambda)$, is finite as long as $\lambda < \infty$. If the probability of an order arriving on the market never equals zero, despite a very high fee charged, the probability that all $N(\lambda)$ trades are at the ask is strictly positive, hence the limit bid order will never execute. It is not necessary that all trades have to occur at the ask. The change in the ask price upwards due to the large number of trades at the ask, causes the market maker also to revise his bid quotes upwards as the result of inventory costs or update of believes. This may result in bid quotes exceeding the limit bid price even if it has initially been the best available bid price. In the same manner other investors can place limit orders that offer a higher price. Hence the probability of execution is always smaller than one, even if the limit bid is very close to the ask, while for $p_b^L \geq p_a$ the order executes with certainty. We have found a probability jump in execution at $p = p_a$:

$$\lim_{p_b^L \to p_a} \phi(p_b^L, \lambda) < 1. \tag{4.243}$$

If more orders arrive at the market, i.e. if λ is increased, the probability that the order will not execute is reduced as can easily be seen from the above example. For $\lambda < \lambda'$ we have for all $p_b^L < p_a$:

$$\phi(p_b^L, \lambda) < \phi(p_b^L, \lambda'). \tag{4.244}$$

[77] Lo et al. (2000) show empirically that the time until a limit order is executed reacts very sensitive to the limit price, whereas other variables, like order size, are of minor importance.

When $\lambda \to \infty$ the probability of non-execution of the limit order goes to zero, i.e. for all $p_b^L < p_a$ we find

$$\lim_{\lambda \to \infty} \phi(p_b^L, \lambda) = 1 \tag{4.245}$$

and the jump in probability vanishes. A continuous trade, i.e. $\lambda = \infty$, requires the investors to revise their portfolios at every instant of time. This is a reasonable assumption as long as trading is costless. If trading imposes costs, investors will not revise their portfolios too often as this would exceed their benefits. Hence in the presence of trading costs other than the fee charged by the market maker, like costs for submitting orders, we find that $\lambda < \infty$ and hence a probability jump at $p = p_a$.

The investor would never submit a limit bid order such that $p_b^L > p_a$ as has been stated above. The expected utility of submitting such an order would be falling and its maximum would be at $p_b^L = p_a$, i.e. by submitting a market order.

Assuming the utility function to be continuous, the expected utility will make a jump downwards by lowering p_b^L below p_a as a result of the jump in the execution probability. Lowering p_b^L further, increases the profits from trading at the stated prices, but also the probability of execution reduces as the fee charged increases. Depending on the slope of the utility curves and the sensitivity of investors to changes in the fee the expected utility can either increase or decrease, also a maximum or minimum at some point is possible. In general no shape can be predicted, COHEN ET AL. (1981) assume that the probability of execution at first reduces only slowly, increasing expected utility, and then it decreases faster, decreasing also expected utility. There exists some point where the expected utility reaches its maximum.

If p_b^L reaches p_b, i.e. the best available bid price, the expected utility jumps again downwards. If submitting a limit order at the same price the order flow has to shared, hence the probability of execution is reduced by its share in the order flow. If the limit order is submitted at prices even lower as p_b at first the limit orders with a better bid price have to be executed before this order is executed. By lowering the bid price further and further, an increasing number of limit orders has to be

executed before. Depending on the utility function, the sensitivity of investors to changes in fees and the distribution of other limit bid orders the expected utility may be falling or increasing.

Figure 4.14 presents an example how the expected utility may look like. In the first panel the investor will choose to submit a limit bid order at $p_b^L = p_b'$. The spread is reduced from s to s'. If the expected utility at $p_b^L = p_a$ is higher, as in the second panel, the investor would submit a market buy order.[78]

The jump in expected utility at $p = p_a$ prevents the spread from converging to zero with an increased number of limit orders submitted. There always exists a price below p_a such that the expected utility from submitting a market order is higher, hence the spread will never be below a certain threshold. The spread that can be achieved depends on the size of the probability jump at p_a. With a low λ this jump will be larger, confirming the empirical finding that more frequently traded assets have smaller spreads than less actively traded assets. For this result no adverse selection costs are needed as in the previous models to explain this result.

Although this analysis of COHEN ET AL. (1981) is very intuitive, it faces several problems for the analysis of the behavior of spreads. For determination of the probability of execution dynamic aspects have to be taken into account. In every N trades new limit orders can be submitted, reducing the execution probability of existing limit orders by offering a more favorable price. Also adverse selection costs may become a severe problem if the limit order cannot be withdrawn. The next section addresses some of these aspects.

4.5.2 Price movements with limit order trading

The order submission strategy does not only depend on unexecuted limit orders in the market, but also on the expectations of future order submissions during the

[78] In all cases the strategy to submit an order has to be compared with the expected utility of submitting no order at all. If the expected utility from submitting no order is higher, the investor will not submit an order as he has no obligation to do so.

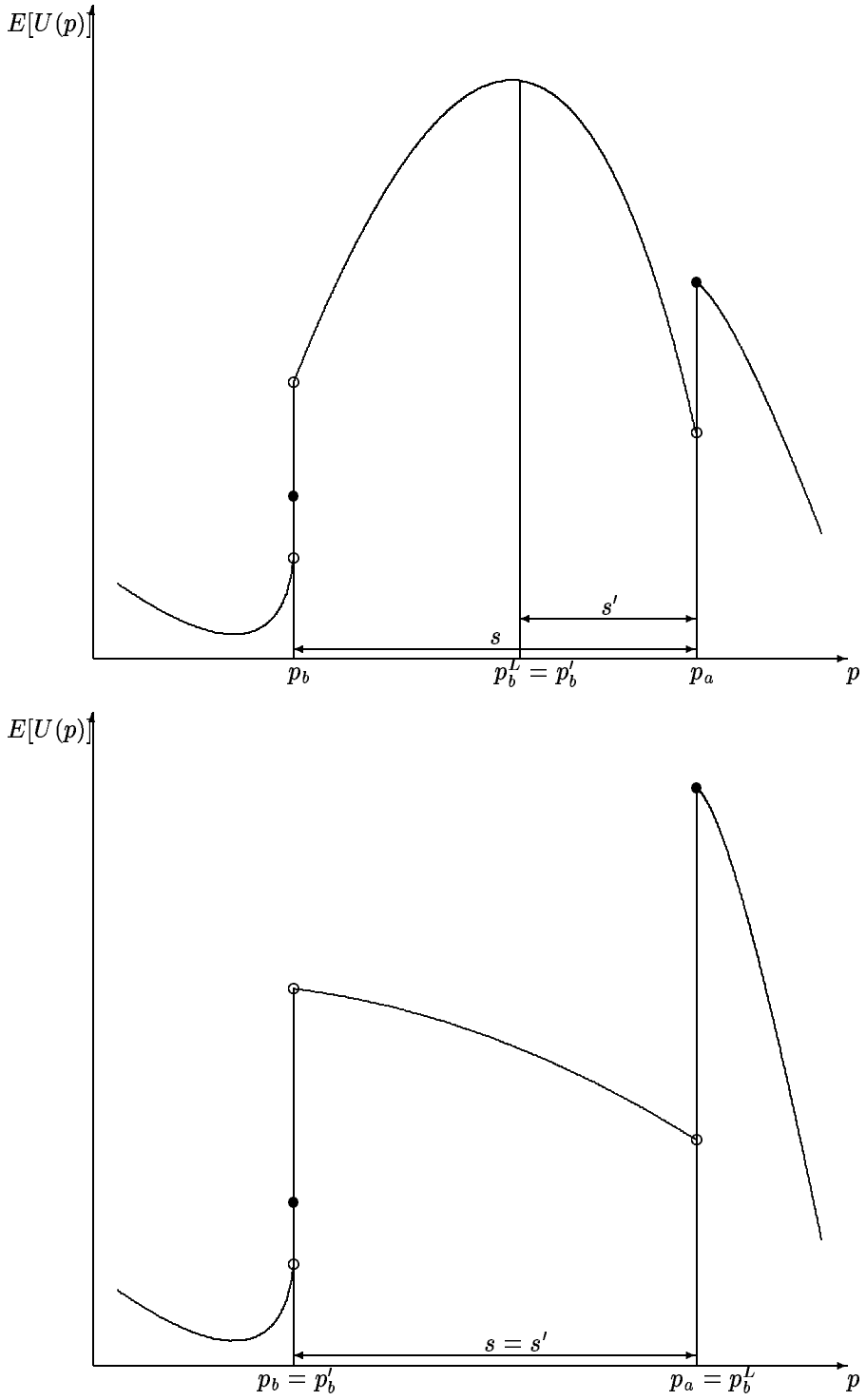


Figure 4.14: The placement of limit orders

remaining trading rounds. Limit orders submitted at more favorable prices will have priority and therefore reduce the probability of execution. But also limit orders submitted at the other side of the trade or at less favorable prices influence the execution probability, because submitted orders are no market orders that have to be executed against a limit order. Most models of limit order trading presented in the literature thus far are static and therefore do not allow to model these dynamic aspects. The first to model the order submission strategy in a dynamic environment is the recent work by PARLOUR (1998). She models not only the order submission strategy in more detail than COHEN ET AL. (1981), but also its implications for the movement of prices.

The time structure of the model consists of two periods. Every investor has an initial endowment of the asset which is optimal for him. In the first period he can choose to trade the asset at given bid and ask prices, p_b and p_a, respectively. He can buy the asset and finance this by reducing his consumption in period 1, C_1, according to the price he is charged for the asset, or he can sell the asset and consume the amount received. In period 2 the asset is liquidated at the fundamental value p^*, which is known to all market participants, hence there are no informational asymmetries, adverse selection or even uncertainty about this value. All market participants have to reverse their trades of period 1 by reducing their consumption in period 2, C_2, by p^* if they have sold the asset or to increase consumption by the same amount if they have bought the asset before. Hence the asset is a mean to delay or advance consumption between the two time periods. The trade size is assumed to be fixed.

The investors are assumed to be risk neutral with a utility function

$$U(C_1, C_2) = C_1 + \beta C_2, \tag{4.246}$$

where $\beta \geq 0$ is a parameter that denotes the preferences for consumption in periods 1 and 2.[79] For $\beta < 1$ consumption in period 1 is preferred and for $\beta > 1$ consumption in period 2. The more β deviates from 1 the more urgent is the need for trading

[79] In terms of microeconomics it is the rate of substitution between consumption in these two periods as the slope of the indifference curve can easily be shown to be $\frac{\partial C_1}{\partial C_2} = -\beta$.

in the respective period as we will see. If β is close to 1, investors are very patient with trading.

Investors differ only in their value for β, which we assume them to know. We further assume that they know the distribution F of the β's of the other investors.

Trading takes place only in period 1 in a fixed number of $T \geq 1$ trading rounds. In each trading round an investor is randomly chosen to arrive on the market. Investors are chosen according to the distribution function F.[80] The chosen investor has to decide whether he wants to submit a market buy, limit buy, limit sell, market sell order or not to trade at all.

If an investor submits a limit order he is not free to choose his limit price. The limit price is assumed to be fixed at p_a for a limit ask order and at p_b for a limit bid order, which are the current bid and ask prices. This strong assumption is made to isolate the effect limit order submissions have on the behavior of investors. The absence of any inventory or adverse selection costs prevent the bid and ask prices from having any dynamics due to the behavior of a market or match maker. But it also restricts the price competition of limit orders. The orders submitted in this way are executed with arriving market orders according to time priority, i.e. all orders that were submitted in earlier trading rounds are executed first. It is obvious that the more limit orders are already unexecuted in the limit order book, the less likely is the execution of a newly submitted limit order. With a finite number of trading rounds the probability of execution will always be lower than one.

An investor submitting a market sell order receives the amount of p_b in period 1 and has to repurchase the asset in period 2 at p^*. When submitting a limit sell order he receives the amount p_a in period 1 and has to repurchase the asset in period 2 also at p^*, but only if his limit order executes, what happens with probability π^s. When submitting a market buy order he pays p_a in period 1 and receives p^* in period 2. With a limit buy order he pays only p_b in period 1 and also receives p^* in period

[80] We assume that there are much more investors than trading rounds, so that we do not have to care about an investor to be chosen twice to trade.

2, provided that his limit order executes, what has a probability of π^b. If the limit orders do not execute, he will receive no payment and also has to make no payments, the same situation if he decides not to trade at all. The expected utility is therewith given by

$$E[U(C_1, C_2)] = \begin{cases} p_b - \beta p^* & \text{market sell order} \\ \pi^s(p_a - \beta p^*) & \text{limit sell order} \\ 0 & \text{no order} \\ \pi^b(-p_b + \beta p^*) & \text{limit buy order} \\ -p_a + \beta p^* & \text{market buy order} \end{cases} . \tag{4.247}$$

The optimal strategy is to maximize (4.247). Comparing market and limit sell orders, gives the condition for preferring a market sell order if $p_b - \beta p^* > \pi^s(p_a - \beta p^*)$, transforming into

$$\beta < \beta_M^s = \frac{p_a}{p^*} - \frac{p_a - p_b}{p^*(1 - \pi^s)} < \frac{p_b}{p^*}. \tag{4.248}$$

If β_M^s turns out to be negative then we have to set $\beta_M^s = 0$. Hence if $\beta < \beta_M^s$ submitting a market sell order is preferred to submitting a limit sell order. Comparing a limit sell order with not trading gives $\pi^s(p_a - \beta p^*) > 0$:

$$\beta < \beta_L^s = \frac{p_a}{p^*}, \tag{4.249}$$

if $\beta < \beta_L^s$ submitting a limit order will be preferred to not trading. The submission of a limit buy order is preferred to not trading if $\pi^b(-p_b + \beta p^*) > 0$:

$$\beta > \beta_L^b = \frac{p_b}{p^*}. \tag{4.250}$$

If $\beta < \beta_L^b$ then not to submit an order is preferred to submitting a limit buy order. For preferring a limit buy to a market buy order we need $\pi^b(-p_b + \beta p^*) > -p_a + \beta p^*$:

$$\beta < \beta_M^b = \frac{p_b}{p^*} + \frac{p_a - p_b}{p^*(1 - \pi^b)}. \tag{4.251}$$

The final comparison that has to be made is for comparing a limit sell and a limit buy order as they are both preferred to the other alternatives for $\beta_L^b < \beta < \beta_L^s$. The condition that a limit sell order is preferred, $\pi^s(p_a - \beta p^*) > \pi^b(-p_b + \beta p^*)$, solves to

$$\beta < \beta_L^{bs} = \frac{p_a}{p^*} - \frac{\pi^b}{\pi^b + \pi^s} \frac{p_a - p_b}{p^*}. \tag{4.252}$$

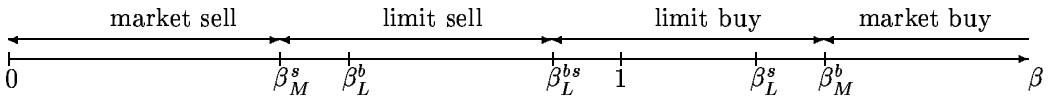


Figure 4.15: Order submission strategies

By the transitivity of preferences all alternatives can now be clearly ordered. Figure 4.15 illustrates the order submission strategy. For extreme values of β investors prefer to submit market orders to ensure consumption in their preferred period. If β is close enough to 1, the need to consume in the preferred period is less important than the possible gain from receiving a more favorable price, hence limit orders are submitted. In all cases the alternative not to trade is dominated.

The probability of execution of a limit order, π^b and π^s, plays a central role like in COHEN ET AL. (1981). Not only the past order submissions, which are known to all investors from the open order book, are of importance, but also the future behavior of the investors. As the distribution of β is known, we could in general use (4.248) - (4.252) to determine the probabilities for the different order types to be submitted in the remaining trading rounds. Unfortunately the important parameters β_M^s and β_M^b depend themselves on these probabilities. Therefore PARLOUR (1998) proposes an indirect approach to address this problem.

Let b_t^B and b_t^A denote the number of unexecuted limit buy and sell orders in the market. As the execution probability depends on these parameters, also β_M^s and β_M^b will depend on them. PARLOUR (1998, pp. 809 ff.) gives a detailed proof that

(4.253)
$$\begin{aligned} \beta_M^s(b_t^B, b_t^A) &\geq \beta_M^s(b_t^B, b_t^A - 1), \\ \beta_M^s(b_t^B, b_t^A) &\geq \beta_M^s(b_t^B + 1, b_t^A), \\ \beta_M^s(b_t^B, b_t^A) &\leq \beta_M^s(b_t^B + 1, b_t^A + 1), \end{aligned}$$

(4.254)
$$\begin{aligned} \beta_M^b(b_t^B, b_t^A) &\leq \beta_M^b(b_t^B - 1, b_t^A), \\ \beta_M^b(b_t^B, b_t^A) &\leq \beta_M^b(b_t^B, b_t^A + 1), \\ \beta_M^b(b_t^B, b_t^A) &\geq \beta_M^b(b_t^B + 1, b_t^A - 1). \end{aligned}$$

By noting that

$$\begin{aligned} \frac{\partial \beta_M^s}{\partial \pi^s} &\leq 0, \\ \frac{\partial \beta_M^b}{\partial \pi^b} &\geq 0 \end{aligned} \tag{4.255}$$

as can easily be seen by differentiating (4.248) and (4.251), we can relate the execution probabilities to the number of outstanding limit orders. Surprisingly the execution probabilities depend on both sides of the limit order book and not only on the side at which the order executes.

For observing a transaction at the ask price a market buy order has to be submitted. A market buy order is submitted if the investor chosen has a β such that $\beta > \beta_M^s$, which will be observed with probability $1 - F(\beta_M^s(b_{t+1}^B, b_{t+1}^A))$. With p_t denoting the price observed in trading round t and o_t the order type at time t we get

$$\text{Prob}(p_{t+1} = p_a | o_t) = 1 - F(\beta_M^s(b_{t+1}^B, b_{t+1}^A) | o_t). \tag{4.256}$$

If in trading round t also a market order has been submitted, we find that $b_{t+1}^B = b_t^B$ and $b_{t+1}^A = b_t^A - 1$. With (4.253) we get

$$\beta_M^s(b_{t+1}^B, b_{t+1}^A) = \beta_M^s(b_t^B, b_t^A - 1) \leq \beta_M^s(b_t^B, b_t^A). \tag{4.257}$$

If the order has been a market sell order we have $b_{t+1}^B = b_t^B - 1$ and $b_{t+1}^A = b_t^A$. Again with (4.253) we get

$$\beta_M^s(b_{t+1}^B, b_{t+1}^A) = \beta_M^s(b_t^B - 1, b_t^A) \geq \beta_M^s(b_t^B, b_t^A). \tag{4.258}$$

Obviously the expression $\beta_M^s(b_{t+1}^B, b_{t+1}^A)$ in equation (4.257) is smaller than in equation (4.258). By the monotonicity of the distribution function we have

$$F(\beta_M^s(b_{t+1}^B, b_{t+1}^A) | p_t = p_a) \leq F(\beta_M^s(b_{t+1}^B, b_{t+1}^A) | p_t = p_b). \tag{4.259}$$

Inserting into (4.256) gives the results that

$$\text{Prob}(p_{t+1} = p_a | p_t = p_a) \geq \text{Prob}(p_{t+1} = p_a | p_t = p_b). \tag{4.260}$$

$$\text{Prob}(p_{t+1} = p_b | p_t = p_a) \leq \text{Prob}(p_{t+1} = p_b | p_t = p_b), \tag{4.261}$$

It turns out that it is more likely to observe subsequent trades both at the bid or at the ask rather than a change of the side of the trade. As in section 4.4.4 we now can calculate the covariance of subsequent price changes. It will remain negative, but as the probability that the price change equals zero is increased owing to (4.260) and (4.261), we get

$$Cov(\Delta p_{t+1}, \Delta p_t) \geq -\frac{s^2}{4}. \tag{4.262}$$

We therewith found first that the strategies to submit limit orders does not only depend on the number of unexecuted limit orders on the same side of the trade, but also on the number of unexecuted limit orders on the other side of the trade. Secondly, we found that it is more likely to observe two subsequent trades on the same side of the trade than on different sides. Hence the first order autocorrelation of transaction prices will be larger than without the possibility to submit limit orders, where it has been found to be -.5.

4.5.3 Informational efficiency with limit order trading

A formal model of the informational efficiency of prices allowing for limit order trading has been proposed by BROWN AND ZHANG (1997). As their model returns to the auction models of chapter 4.2, we do not derive their results explicitly, but give the intuition behind their findings.

If only market orders are allowed to be submitted to the market, informed investors face the risk that due to a large order imbalance arising from uninformed investors, the price is less favorable than expected, they may even make losses. Hence informed investors will trade much less aggressive and prices do not reveal so much information.

If informed investors and risk averse hedgers are allowed to submit limit orders to protect themselves from a too unfavorable shift in prices, both groups will trade more actively. As informed investors can rule out losses from trading, uninformed

investors still face the risk of making a loss as they do not have access to the same information. Therefore informed investors will trade much more aggressively than uninformed investors to exploit their informational advantage.

If all informed investors had perfect knowledge of the fundamental value, as assumed in chapter 4.2, the match maker would observe a bundling of limit orders at this value and would be able to deduct the true value with certainty from observing the limit order flow. By quoting this price the match maker would hinder the informed investors from making any profits. As there are no possible profits from being informed, there are no incentives to become informed, hence with the results of Grossman and Stiglitz (1980) no equilibrium would exist if information acquisition is costly.

To avoid such a situation we can assume that informed investors do not have perfect knowledge of the fundamental value, but observe only a noisy signal. These different observations of the information gives rise to different limit prices and the match maker cannot directly deduct the fundamental value of the asset.

The more aggressive trades of informed investors, nevertheless, make it easier to deduct the fundamental value of the asset and with the possibility to submit limit orders, prices become more efficient than if only market orders were allowed to be submitted. On the other hand, the liquidity of the market is reduced, imposing higher trading costs on uninformed investors. Which effect dominates, depends on the exact specification of the parameters, like the risk of the asset, the risk aversion of uninformed investors and so forth. In a similar model this result is confirmed by Easley and O'Hara (1991).

4.6 Analyzing markets with multiple assets

The models considered thus far assumed the existence of only a single asset. This strong abstraction from reality gave several important insights into the behavior of

prices and the emergence of quotes in dealer markets. The importance of covariances between assets has been pointed out in the literature not only for asset pricing, but also for the portfolio choices of investors (see also appendix B). We can therefore expect to gain additional insights by extending the analysis to multi-asset markets.[81]

Unfortunately, contributions to multi-asset markets are not frequently found, only three contributions exists to the authors best knowledge that use the framework provided in this chapter. One article by CABALLE AND KRISHNAN (1994) focuses on auction markets, while HAGERTY (1991) and GEHRIG AND JACKSON (1998) investigate the price setting behavior of a monopolistic market maker.

4.6.1 Auction markets with multiple assets

Using the framework developed in chapter 4.2.2, only restricted to the case of a single trading round, CABALLE AND KRISHNAN (1994) extend this model to the case of investors trading in multiple assets.

Each investor receives a noisy signal of the liquidation values v, such that[82]

$$\begin{aligned} s_i &= v + \varepsilon_i, \\ \varepsilon_i &\sim N(0, \Sigma_\varepsilon), \end{aligned} \tag{4.263}$$

where $s_i = (s_i^1, \ldots, s_i^L)$ denotes the vector of signals investor i receives for each of the L assets, $v = (v^1, \ldots, v^L)$ the vector of the liquidation values, $\varepsilon_i = (\varepsilon_i^1, \ldots, \varepsilon_i^L)$ the vector of noise terms as received by investor i and Σ_ε the covariance matrix of

[81] CHORDIA ET AL. (2000) give empirical evidence that the spread is affected by asset specific factors as well as common factors. Although they do not provide a more detailed analysis, we can expect that, depending on the correlation of two assets, the influence of common factors depends also on the correlation between assets.

[82] With the exception of this assumption, where we assumed perfect knowledge of the liquidation value, the assumptions are identical to those in chapter 4.2.2. This modification is necessary to rule out that the match maker can learn the information perfectly through aggregating information on different assets. As repeatedly pointed out, in this case there would exist no equilibrium in information acquisition as long as information acquisition is costly. It has to be noticed that here all variables represent vectors or matrices of the considered assets.

noise terms. The liquidation value is multivariate normally distributed with mean $p_0 = (p_0^1, \ldots, p_0^L)$ and covariance matrix Σ_0:

$$v \sim N(p_0, \Sigma_0). \tag{4.264}$$

The imbalance of orders from uninformed investors, $u = (u^1, \ldots, u^L)$, is normally distributed with mean zero and covariance matrix Σ_u:

$$u \sim N(0, \Sigma_u). \tag{4.265}$$

The match maker sets prices such that with $p = (p^1, \ldots, p^L)$ denoting the vector of prices and the aggregate order flow from informed investors $x = (x^1, \ldots, x^L)$, $x^j = \sum_{i=1}^{M} x_i^j$, where x_i^j represents the order of informed investor i for asset j:

$$p = E[v|x + u]. \tag{4.266}$$

The profits from trading for informed investors are the sum of the profits made from trading each asset:

$$\pi_i = (v - p)' x_i, \tag{4.267}$$

where $x_i = (x_i^1, \ldots, x_i^L)$. As in chapter 4.2.2, the analysis is restricted to a linear equilibrium:

$$p = \mu + \Lambda(x + u), \tag{4.268}$$

$$x_i = \alpha + B(s_i - p_0), \tag{4.269}$$

where $\mu = (\mu_1, \ldots, \mu_L)'$ and $\alpha = (\alpha_1, \ldots, \alpha_L)'$ are vectors of constants and $\Lambda = (\lambda^{11}, \ldots, \lambda^{1L}; \ldots; \lambda^{L1}, \ldots, \lambda^{LL})$ and $B = (\beta^{11}, \ldots, \beta^{1L}; \ldots; \beta^{L1}, \ldots, \beta^{LL})$ are positive definite and nonsingular matrices of constants.

Informed investors are assumed to be risk neutral and hence maximize expected profits given their information, which are given by

$$E[\pi_i|s_i] = E\left[(v - p)' x_i | s_i\right] \tag{4.270}$$

Following steps similar to those presented in chapter 4.2.2 and as presented in more detail by CABALLE AND KRISHNAN (1994, pp. 701 f.), it turns out that

$$p = p_0 + \frac{\sqrt{M}}{2}\Gamma x, \tag{4.271}$$

$$x = \Theta(s_i - p_0), \tag{4.272}$$

where

$$\begin{aligned} \Gamma &= \Sigma_u^{-\frac{1}{2}} H^{\frac{1}{2}} \Sigma_u^{-\frac{1}{2}}, \\ H &= \Sigma_u^{\frac{1}{2}} G \Sigma_u^{\frac{1}{2}} \\ G &= \left[\Sigma_\varepsilon^{-1} + \frac{M-1}{2}\Sigma_\varepsilon^{-1}\Sigma_s\Sigma_\varepsilon^{-1}\right]^{-1} \\ &\quad - \left[\frac{2}{M-1}\Sigma_s^{-1} + 2\Sigma_\varepsilon^{-1} + \frac{M-1}{2}\Sigma_\varepsilon^{-1}\Sigma_s\Sigma_\varepsilon^{-1}\right]^{-1}, \\ \Sigma_s &= \Sigma_0 + \Sigma_\varepsilon, \\ \Theta &= \frac{1}{\sqrt{M}\Gamma^{-1}\left[I + \frac{M-1}{2}\Sigma_s\Sigma_\varepsilon^{-1}\right]^{-1}}. \end{aligned} \tag{4.273}$$

Although an analytical solution exists, it is very difficult to interpret due to its complexity. However, some general properties can be derived. It can be shown that Γ is positive definite and symmetric. We can interpret this property as follows: the order flow of the pth asset affects the price of the qth asset in the same degree as the order flow of the qth asset influences the price of the pth asset. This means that the informativeness of trades in the other asset is equal for all assets.[83]

Further insights may be gained by analyzing the correlation structure in more detail, i.e. how different covariances in error terms or liquidity trade imbalances influence the covariance of the observed prices. The complexity of the result makes such an analysis very difficult to conduct in general. It should be noted that as

$$Var[v|p] = \Sigma_0 - M\left[2\Sigma_\varepsilon^{-1} + (M-1)\Sigma_\varepsilon^{-1}\Sigma_s\Sigma_\varepsilon^{-1}\right]^{-1}, \tag{4.274}$$

the beliefs after observing prices can exhibit a completely different correlation structure than the initial beliefs. Further properties of this model are very difficult to

[83] CABALLE AND KRISHNAN (1994, p.700) report that this result can be confirmed in empirical investigations.

derive analytically. Here a numerical analysis and simulations may bring further insights, as can be expected from extending to model to allow for more than only a single trading round.

4.6.2 Dealer markets with multiple assets

Let us now address dealer rather than auction markets. We assume a market with $L > 1$ different risky assets. Each asset is assigned to a single market maker, who is granted a monopoly to act as market maker for this asset. This market structure can be found on the NYSE for example, where these market makers are called specialists. We allow each specialist to make the market for more than a single asset, i.e. be specialist in more than one asset.[84] We have $K \leq L$ specialists, which are assumed to be risk neutral, i.e. they face no inventory costs. It is further assumed that no investor has private information. With this assumption we avoid the problem of adverse selection costs. The aim of these assumptions is to concentrate on the effect the execution of monopoly power has in this model. In a market as described here, the spread of competitive market makers would be zero as we know from sections 4.3 and 4.4.

Following Gehrig and Jackson (1998) we suppose two groups of investors, indexed $i = \{1, 2\}$. Non-optimal endowments in their portfolio holdings induce them to trade, i.e. trading takes place as the result of a liquidity event. One group buys an asset and the other sells it. Let the endowment of investors in group i of asset j be denoted e^i_j and $e^i = (e^i_1, \ldots, e^i_L)$ is the vector of endowments. The total supply is $e = e^1 + e^2$.

Specialist k controls a subset L_k of the L assets. Specialists set prices such that the market clears, i.e. demand equals supply. Equilibrium prices exist if we assume the specialist to know the endowments of the investors and all other relevant variables.

[84] Hagerty (1991) allows every specialist only to make the market in one asset, so that the framework considered here is a generalization of her setting.

The risky assets are traded in a single round of trading, after the trade has occurred they are liquidated at the fundamental value. The fundamental value of asset j is denoted v_j and has a mean of μ_j and a covariance matrix Σ, which we assume to be positive definite and non singular.

A risk averse investor will determine the demand such that he maximizes his expected utility of terminal wealth, which is given by[85]

$$W^i = \sum_{j=1}^{L} \left(x_j^i v_j + p_{b_j} \max(e_j^i - x_j^i, 0) + p_{a_j} \min(e_j^i - x_j^i, 0)) \right), \tag{4.275}$$

where x_j^i denotes the demand of investor i for asset j, p_{b_j} the bid and p_{a_j} the ask price of asset j. The final terms denote the change in money holding as a result of trading and the first term the return from holding the asset. With a competitive market maker we have shown earlier that in this environment $p_{b_j} = p_{a_j} \equiv p_j^c$. Inserting this we get

$$W^i = \sum_{j=1}^{L} \left(x_j^i v_j + p_j^c (e_j^i - x_j^i) \right), \tag{4.276}$$

$$E[W^i] = \sum_{j=1}^{L} \left(x_j^i \mu_j + p_j^c (e_j^i - x_j^i) \right) = x^i{}' \mu + p^c{}' (e^i - x^i), \tag{4.277}$$

$$\begin{aligned} Var[W^i] &= \sum_{j=1}^{L} (x_j^i)^2 \Sigma_{jj} + \sum_{j=1}^{L} \sum_{k=1, k \neq j}^{L} x_j^i x_k^i \Sigma_{jk} \\ &= x^i{}' \Sigma x^i. \end{aligned} \tag{4.278}$$

The expected utility is then given by

$$E[U(W^i)] = U(E[W^i] - \frac{1}{2} z Var[W^i]). \tag{4.279}$$

Inserting and differentiating gives the first order condition for a maximum:

$$\mu - p^c - z \Sigma x^i = 0. \tag{4.280}$$

Solving for the optimal demand we obtain

$$x^i = \frac{1}{z} \Sigma^{-1} (\mu - p^c). \tag{4.281}$$

[85] As done repeatedly before, we focus only on the portfolio imbalance rather than the entire wealth. The portfolio imbalance here is the endowment of the investors.

The second order condition $-z\Sigma < 0$ is fulfilled with the assumption of a positive definite covariance matrix. The market clearing condition $x^1 + x^2 = e$ gives us the competitive price as

$$p^c = \mu - \frac{1}{2} z\Sigma e. \tag{4.282}$$

Let $\theta^i = x^i - e^i$ denote the vector of trades investor i conducts. If the specialist has market power he is able to charge different prices for the two groups of investors, denoted p^i. Therewith (4.281) becomes

$$x^i = \frac{1}{z}\Sigma^{-1}(\mu - p^i). \tag{4.283}$$

Market clearing requires that $\theta^1 + \theta^2 = 0$:

$$\begin{aligned} 0 &= \theta^1 + \theta^2 = x^1 + x^2 - e \\ &= \frac{1}{z}\Sigma^{-1}(2\mu - p^1 - p^2) - e. \end{aligned} \tag{4.284}$$

Solving (4.282) for e and inserting into (4.284) gives

$$\frac{1}{z}\Sigma^{-1}(2\mu - p^1 - p^2) - \frac{2}{z}\Sigma^{-1}(\mu - p^c) = \frac{1}{z}\Sigma^{-1}(2p^c - p^1 - p^2) = 0, \tag{4.285}$$

what requires[86]

$$p^1 - p^c = p^c - p^2. \tag{4.286}$$

The specialist, assumed to be risk neutral, maximizes his total profits. With R_j^l denoting the revenues from asset j for specialist l we get with the market clearing condition $\theta^1 + \theta^2 = 0$ and (4.286):

$$R_j^l = \theta_j^1 p_j^1 + \theta_j^2 p_j^2 = \theta_j^1 p_j^1 - \theta_j^1(2p_j^c - p_j^1) = 2\theta_j^1(p_j^1 - p_j^c). \tag{4.287}$$

The total revenues of specialist l are given by

$$R^l = \sum_{j \epsilon L_l} R_j^l = 2\sum_{j \epsilon L_l} \theta_j^1(p_j^1 - p_j^c) = 2\sum_{j \epsilon L_l}(x_j^1 - e_j^1)(p_j^1 - p_j^c). \tag{4.288}$$

[86] It can be shown that only a symmetric solution is optimal. This result is the consequence of the market maker having the same market power on both sides of the market.

Differentiating with respect to p_j^1 for all $j \in L_l$ we get with (4.283) the first order conditions for each $j \in L_l$

$$\frac{1}{z}\left(\Sigma^{-1}\right)_j'(\mu - p^1) - e_j^1 - \frac{1}{z}\sum_{i \epsilon L_l} \Sigma_{ij}^{-1}(p_i^1 - p_i^c) = 0, \tag{4.289}$$

where $\left(\Sigma^{-1}\right)_j$ denotes that jth row of Σ^{-1} and Σ_{ij}^{-1} the (i,j)th element of Σ^{-1}. Defining

$$A_{ij} = \begin{cases} 2\Sigma_{ij}^{-1} & \text{if } i = j \text{ or } i,j \in L_l \\ \Sigma_{ij}^{-1} & \text{else} \end{cases}, \tag{4.290}$$

we can rewrite (4.289) in vector form as

$$\frac{1}{z}\Sigma^{-1}(\mu - p^c) - e^1 - \frac{1}{z}A(p^1 - p^c) = 0, \tag{4.291}$$

where from (4.281) and (4.283)

$$\mu - p^c = z\Sigma x^1 = z\Sigma\frac{1}{z}\Sigma^{-1}(\mu - p^1) = \mu - p^1. \tag{4.292}$$

Solving (4.291) for $p^1 - p^c$ we get

$$p^1 - p^c = A^{-1}\left(\Sigma^{-1}(\mu - p^c) - ze^1\right). \tag{4.293}$$

As from (4.282) we obtain

$$e = \frac{2}{z}\Sigma^{-1}(\mu - p^c) \tag{4.294}$$

we find

$$\begin{aligned} p^1 - p^c &= \frac{1}{2}zA^{-1}\left(\frac{2}{z}\Sigma^{-1}(\mu - p^c) - 2e^1\right) \\ &= \frac{1}{2}zA^{-1}(e - 2e^1) \\ &= \frac{1}{2}zA^{-1}(e^2 - e^1). \end{aligned} \tag{4.295}$$

If an investor of group 1 buys the asset, i.e. $e_j^1 < e_j^2$, we find that $p_j^1 > p_j^c > p_j^2$ and we can interpret p_j^1 as the ask price and p_j^2 as the bid price. If an investor of group 2 buys the asset the relations change. The spread is given by

$$s = |p^1 - p^2| = z|A^{-1}(e^2 - e^1)| > 0. \tag{4.296}$$

The bid and ask prices are symmetric around p^c as can be seen from (4.286), because the market maker has the same market power on both sides of the trade. The spread is increasing in the risk aversion of the investors. With a higher risk aversion the demand reacts less elastic to price changes, because the risk of holding a non-optimal portfolio has a larger influence on the expected utility than trading costs. In the same manner differences in endowment, i.e. the portfolio imbalance, increases the spread. The exact influence on the spread depends on the covariances of the assets, that form A. If two assets are very similar, i.e. their correlation is close to 1, the investor can trade the other asset instead and receive only a small reduction in expected utility, because assets are close substitutes. This indirect competition between assets (if they are not assigned to the same market maker) forces the specialists to compete against each other, what reduces their market power and hence the spread.

Using this result, GEHRIG AND JACKSON (1998) derive some further properties by analyzing the correlation structure of the assets and the resulting spreads in more detail. Comparing the situation with two specialists each being assigned to one asset and a single specialist being responsible for both assets, the endowments of the investors have to be considered. If the two groups of investors are both well engaged in the asset market and only want to rebalance their portfolios, it turns out that a single market maker quotes a lower spread for positively correlated assets, while two specialists would give lower spreads for negatively correlated assets.

Investors rebalancing their portfolio cannot gain significant expected utility from trading two assets with positive correlation. Hence the joint profit maximization of a single specialist limits the execution of market power more than with two independent specialists and gives rise to lower spreads. With negatively correlated assets this relationships reverses. Trading negatively correlated assets would give an investor a larger increase in expected utility, hence a single specialist can execute his market power much more than two independent specialists.

If the investors, however, want to change their overall engagement in the asset market without changing the composition of their portfolios, the relation exactly reverses.

For positively correlated assets two specialists would give lower spreads and a single specialist for negatively correlated assets.

This result can easily be understood with the indirect competition between assets. Suppose it is optimal for an investor to trade asset i. However, if the costs of trading this asset are too high, he can trade another asset, j, which is positively correlated with asset i without a large loss in expected utility. Hence indirect competition between assets reduces market power and therewith the spread. In contrast, a single specialist for both assets could execute his market power. In the case of a negative correlation between assets, an investor cannot trade asset j instead of asset i without incurring a large loss in expected utility. Independent specialists would have more market power, whereas the joint profit maximization of a single specialist for both assets would limit market power.

The optimal allocation of the responsibilities for assets has not only to take into account the correlation structure of the assets, but also the motives of trading, portfolio rebalancing of well engaged investors or changing the engagement in the entire asset market.

It can be preferable to assign a monopoly in market making for several assets to the same market maker in order to limit the execution of market power, while in other situations indirect competition reduces the spread.

We can summarize these findings by stating that if assets trade as substitutes, the spread is lower if the assets are assigned to different market makers, while for assets trading as complements a single market maker would be preferable.

4.7 Empirical investigations of spread components

In the preceding sections of this chapter we identified two components of the spread, s: *inventory costs*, c_I, and *adverse selection costs*, c_A. Thus far ignored have been

order processing costs, c_O, which arise from maintaining the infrastructure of a market maker, e.g. his computer system and staff, which are fixed costs as well as variable costs arising from conducting a trade, like fees charged by the exchange. We assume that all costs include normal profits as has been determined previously. We then find a final component of the spread, which is mostly neglected in the literature, *excess profits*, c_π. In nontransparent dealer markets, like telephone markets without a dissemination of quotes, e.g. the foreign exchange market, FLOOD ET AL. (1998) show that costs for searching the best quotes account for about a third of the quoted spread. As we only consider exchanges with quote disclosure, we can neglect these search costs.

The spread can therewith be written as the sum of its components:

$$s = c_I + c_A + c_O + c_\pi. \tag{4.297}$$

In this section we will briefly review some contributions on the estimation of these components. No investigation addresses excess profits, for which reason we neglect this spread component in the remainder.

At first we will briefly consider the estimation technique as introduced by STOLL (1989) and then show the results of several empirical investigations.

4.7.1 An estimation technique

There exists a large variety of estimation techniques for the components of the spread. Many of these techniques concentrate on the determination of the adverse selection component and do not further distinguish between inventory and order processing costs. We will here concentrate on the approach developed by STOLL (1989), which explicitly uses the results of market microstructure models as described before.

We will at first investigate the properties of prices if only one of the components is present and finally combine these properties into a single framework used for estimation. The model of STOLL (1989) requires to observe every single trade in an asset to conduct the estimation.[87] We assume for simplicity that the first observed trade has taken place at the bid, results for the first trade at the ask can be derived in the same manner.

At first we investigate the case that only order processing costs are present, i.e. $s = c_O$, which are supposed to be constant. With the fundamental value not changing over time the bid and ask price are the same throughout the sampling period, hence the price change is either zero if the next trade is also at the bid or s if the next trade is at the ask. We assume furthermore that prices are set such that the market clears on average, hence the probability for a trade at the bid and at the ask are both .5. We then have for the change of the transaction price at time t, $\Delta p_t = p_t - p_{t-1}$:

$$\Delta p_t = \begin{cases} s & \text{with probability } .5 \\ 0 & \text{with probability } .5 \end{cases} . \tag{4.298}$$

When only inventory costs are present, i.e. $s = c_I$, we know from the inventory based models of market making that after a trade at the bid both, the bid and ask prices, decrease to adjust for the larger inventory, while the spread is held constant. The linearity of inventory costs in inventory and the symmetry of the price changes implies the price to fall by $.5s$. To see this, notice that the costs for a trade at the bid before and after another trade at the bid has been conducted are from (4.159):

$$C^1 = \frac{1}{2} z\sigma^2 \left(Q^2 + IQ'\right) \tag{4.299}$$

$$C^2 = \frac{1}{2} z\sigma^2 \left(Q^2 + (I+Q')Q'\right) . \tag{4.300}$$

The change in inventory costs for a trade at the bid are

$$\Delta C = C_2 - C_1 = \frac{1}{2} z\sigma^2 Q^2 = \frac{1}{2} s. \tag{4.301}$$

[87] As such data have only been available recently, early contributions developed models using daily data to estimate the components. A major problem many of them address, is to determine the spread as quoted by market makers, as such data also have not readily been available. Such a method has been presented in chapter 4.3.4, but we do not consider these models here in more detail as it has become a standard nowadays to use data on transaction basis.

As the spread does not depend on the inventory, it is obvious that trading costs at the ask reduce by $.5s$. Similar results for trades at the ask can easily be derived in the same manner. The probabilities for a trade at the bid and ask, respectively, are no longer equal. The market maker decreases the ask price and hence the costs for investors trading at the ask, while he increases these costs for a trade at the bid, the probability of receiving such a trade reduces, hence[88]

$$\Delta p_t = \begin{cases} .5s & \text{with probability } .5 < \gamma < 1 \\ -.5s & \text{with probability } 0 < 1-\gamma < .5 \end{cases} . \tag{4.302}$$

A similar argumentation we can use in the presence of solely adverse selection costs, i.e. $s = c_A$. As we know from information based models of market making, prices are set such that they are the expected fundamental value in case a transaction takes place at this side of the market. Assuming adverse selection costs not to change over time and therewith a constant spread, prices also change by $.5s$. The proof of this claim can easily conducted with the model provided in chapter 4.4.1. When assuming the market to clear on average, we find

$$\Delta p_t = \begin{cases} .5s & \text{with probability } .5 \\ -.5s & \text{with probability } .5 \end{cases} . \tag{4.303}$$

Let us now define a $\delta \in [0,1]$ such that $(1-\delta)s$ measures the price change if the side of the trade changes in subsequent trades. Then $-\delta s$ is the price change if the side does not change. Let further $\gamma \in [0,1]$ measure the probability that the side of the trade changes. Using this notation we can rewrite equations (4.298) - (4.303) as

$$\Delta p_t = \begin{cases} (1-\delta)s & \text{with probability } \gamma \\ -\delta s & \text{with probability } 1-\gamma \end{cases} . \tag{4.304}$$

The sequence of possible transaction prices together with their transition probabilities are shown in figure 4.16 for convenience, where the superscripts ^{a} and ^{b},

[88] We saw in chapters 4.3.2 and 4.3.3 that the market spread at which transactions occur, in general will not represent the costs of market making. In the case of a monopolistic market maker the spread will be higher, while in the case of competing market makers, on which we concentrate here, the spread will be lower if there are at least three market makers. To justify the assumption that the market spread corresponds to the reservation spread of a market maker, recall the statement in chapter 4.3.2 that the market spread has a tendency towards the reservation spread. Hence we can use the market spread as an approximation of the reservation spread of a market maker and therewith justify the approach used here.

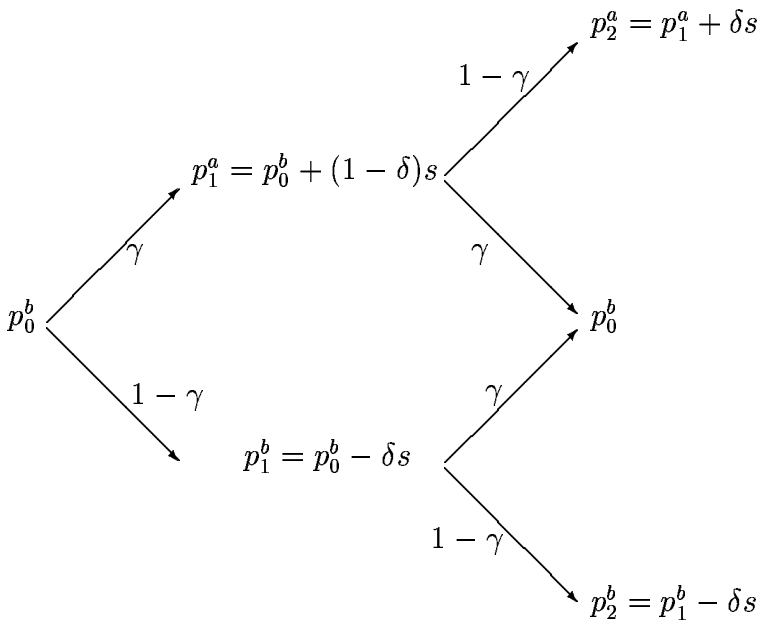


Figure 4.16: Sequences of transaction prices with the initial price at the bid

respectively, denote the ask and bid price. Table 4.1 exhibits the parameter constellations for the different spread components as identified above.

The expected price change between two transactions with the initial trade being at the bid is given by

$$E[\Delta p_t] = (1-\delta)s\gamma + (-\delta s)(1-\gamma) = (\gamma - \delta)s. \tag{4.305}$$

Due to the symmetry, the expected price change with the initial trade at the ask is given by $-(\gamma - \delta)s$ as easily can be shown. We can now define the *realized spread* as the gross revenue made by the market maker from a round trip, i.e. a trade at the bid followed by a trade ask or vice versa. In this case the market maker has the same inventory position and beliefs on the fundamental value of the asset before and after these two trades have taken place and has made a profits from these two transactions that equals the amount of the spread. The realized spread therewith can be calculated as the expected price change after a trade at the bid has taken place, minus the expected price change after a trade at the ask has taken place,

Spread component	δ	γ
Order processing costs	0	$\frac{1}{2}$
Adverse selection costs	$\frac{1}{2}$	$\frac{1}{2}$
Inventory costs	$\frac{1}{2}$	$\frac{1}{2} < \gamma < 1$

Table 4.1: Parameter constellations for the spread components

hence we find the realized spread, s_r, to be

$$s_r = 2(\gamma - \delta)s. \tag{4.306}$$

As we see from table 4.1, the realized spread is zero with only adverse selection costs being present, while in the presence of order processing and inventory costs we find $0 < s_r \leq s$. Hence we can use the realized spread to determine the part of the spread due to adverse selection costs. However, at first we have to determine the parameters γ and δ.

If the probability of the initial trade to take place at the bid is as likely as taking place at the ask, we see that the unconditional expected price change is zero. We then can determine the first order autocovariance of transaction prices with help of the sequences denoted in figure 4.16 to yield

$$\begin{aligned} \rho_T &= Cov[\Delta p_t, \Delta p_{t-1}] \\ &= E[\Delta p_t \Delta p_{t-1}] \\ &= \left[\delta^2(1-2\gamma) - \gamma^2(1-2\delta)\right] s^2. \end{aligned} \tag{4.307}$$

In the same manner we can derive the first order serial covariance of the bid price:[89]

$$\begin{aligned} \rho_Q &= Cov[\Delta p_t^b, \Delta p_{t-1}^b] \\ &= E[\Delta p_t^b \Delta p_{t-1}^b] \\ &= \delta^2(1-2\gamma)s^2. \end{aligned} \tag{4.308}$$

[89] For the ask price it can be shown to yield the same result such that it is of no interest which quoted price is observed.

We can now estimate these covariances from the data sample and determine the parameters γ and δ accordingly. The nonlinearities implied by (4.307) and (4.308) cause severe biases in small sample sizes as reported by BROOKS AND MASSON (1996). Therefore large sample sizes are needed for conducting this estimation.

Having estimated the parameters γ and δ, we can finally determine the spread components. The share of the adverse selection component is given by[90]

$$s_A = \frac{s - s_r}{s} = 1 - 2(\gamma - \delta). \tag{4.309}$$

Using the realized spread we can determine the inventory cost component with the results from table 4.1 by inserting the relevant parameter values into (4.306) as

$$s_I = 2\gamma - 1. \tag{4.310}$$

The remainder is the order processing component:

$$s_O = 1 - 2\delta. \tag{4.311}$$

Besides these estimations several other techniques are used in empirical investigations, in many cases modifications of this estimator. A common approach in most techniques is to decompose the spread into different components and use the properties of each component as a reference point to determine the influence of this component on the entire spread.[91] We will present the results of several empirical investigations in the next section.

4.7.2 Empirical investigations

A large number of empirical investigations addressing the composition of the spread, are reported in the literature. The results of some of these investigations are reported

[90] We here directly define the share a cost component has in the spread as the spread component. We could instead also define the component in absolute values without changing the meaning of the results.

[91] Applying parameter restrictions to the estimation technique presented in HUANG AND STOLL (1997) gives rise to most of the other estimation techniques applied in empirical investigations, including the method in STOLL (1989) presented above. COUGHENOUR AND SHASTRI (1999) give a brief overview of these estimation techniques as well as empirical results presented in the literature.

Authors	Exchange	Period	s_A	s_I	s_O
Glosten and Harris (1988)	NYSE	Dec 1981	.35	.65	
Stoll (1989)	NASDAQ/NMS	Oct-Dec 1984	.43	.10	.47
George et al. (1991)	NASDAQ	1983-1987	.09	.91	
Affleck-Graves et al. (1994)	NYSE/AMEX	Mar/Apr 1985	.59	.29	.12
Affleck-Graves et al. (1994)	NASDAQ	Apr 1985	.35	.24	.41
Schmidt and Treske (1996)	Frankfurt	Feb-Sep 1995	.22	.55	.23
Porter and Weaver (1996)	NYSE	1990	.02	.98	
Porter and Weaver (1996)	AMEX	1990	.16	.84	
Porter and Weaver (1996)	NASDAQ	1990	-.01	1.01	
Huang and Stoll (1997)	NYSE	1992	.10	.29	.61
Madhavan et al. (1997)	NYSE	1990	.40	.60	

Numbers denote average sizes of the components found by the authors for several assets.

Table 4.2: Estimates of the spread components

in table 4.2. The evidence shows large differences between the authors, an observation that cannot only be attributed to the investigation of different assets in different periods of time.

George et al. (1991) point out that the estimator of the spread components in Stoll (1989) is biased for time varying and autocorrelated returns. This may explain the found differences in the estimates between different estimation techniques at least partially. Not for all estimation techniques it is known, whether they give unbiased results if the restrictive assumptions of the underlying models are not exactly met.

Several authors, e.g. Huang and Stoll (1997), use different estimation techniques for the same data set and derive results that differ substantially. Neal and Wheatley (1995) compare two widely applied estimation techniques, those of Glosten and Harris (1988) and George et al. (1991) for various stocks and mutual funds and show large differences in the estimates of adverse selection components, with the estimates of Glosten and Harris (1988) being smaller throughout.

We will therefore have to interpret the results with care. Despite all problems one faces when estimating the spread components, it can be seen that adverse selection

components form a considerable part of the spread. Estimates from many investigations, including several not included in the table, range from about a third to a half of the observed spread. Other investigations, e.g. HANSCH ET AL. (1998), SNELL AND TONKS (1998) and SNELL AND TONKS (1999), find significant evidence of inventory costs, although they do not estimate the share of the spread to be attributed to these costs. HASBROUCK (1988) and STOLL (2000) find evidence on adverse selection as well as inventory costs.

In chapters 4.2.4 and 4.4.1 we showed that a larger share of informed investors increases trading volume and that adverse selection costs increase. We should therefore expect to find a positive relation between trading volume and adverse selection costs. Indeed, GLOSTEN AND HARRIS (1988) report the adverse selection costs and therewith the entire spread, to be increasing in trading volume. A similar result is reported by LAUX (1993) and LIN ET AL. (1995).

The results of Chapter 4.2.2 that adverse selection costs increase and liquidity reduces in the number of informed investors has also been confirmed empirically in CHAKRAVARTY ET AL. (1997).

We may summarize these empirical results shortly by stating that support is found for all spread components to be relevant in markets and adverse selection costs increase with trading volume. However, determining the effect of each component on the total spread is very difficult and subject to the model applied for estimation. Up to now there exists no generally accepted framework for the estimation of spread components and therewith no reliable estimation of the spread components.

4.8 Summary

This chapter presented the most important models in market microstructure theory and some more recent developments.

We showed that immediacy will be supplied by market participants not forced by an exogenous liquidity to trade and that investors benefit from the existence of specialized market participants. It was further shown which market form, auction or dealer market, the providers of immediacy will prefer.

Afterwards we concentrated on the price setting in auction and dealer markets. Using an auction market we found that private information is incorporated gradually into prices, which were shown to be semi-strong efficient from the beginning, whereas they only converge to strong efficiency over time. Strong efficiency of prices is only achieved immediately if informed investors behave perfectly competitive.

In dealer markets inventory costs were shown to give rise to a bid-ask spread. It was found that this spread increases with the risk aversion of market makers, the risk of the asset and the depth of the market. A monopolistic market maker was found to quote spreads larger than competitive market makers by receiving compensation for the risk of an unfavorable inventory shift in the future. Besides the spread is also increasing in the level of demand. If there are more than three competitive market makers, we showed that the market spread is smaller than the spread quoted by a single market maker, but also that in general we will always find a positive spread.

Asymmetric information in dealer markets gives rise to adverse selection costs, what causes a market maker to quote a positive spread. We showed these adverse selection costs to be increasing in the share of informed investors trading. As an increased share of informed investors in general gives rise to higher trading volume, we can expect to find a positive relationship between the spread and trading volume due to higher adverse selection costs.

Including the possibility to submit limit orders as direct competitors to the quotes of market makers can reduce the market spread, but it was shown that it will always remain strictly positive. We further considered the implications of limit order trading on price dynamics.

Generalizing the finding to markets with more than a single asset was shown to

cause severe analytical difficulties. In auction markets the dynamics of asset prices becomes very difficult analyze. For dealer markets we concentrated on the implications of market power. It was shown that for assets trading as substitutes smaller spreads can be achieved by having different market makers for each asset through preventing the execution of market power and foster indirect competition. Whereas for assets trading as complements a single market maker for those assets is optimal.

Empirical investigations were shown in general to support the findings of the theory. But it also became obvious that we face severe difficulties in estimating the different costs components of the spread. The estimation procedure showed up to be very sensitive to the exact specification of the underlying model. Therefore estimates differ widely between authors and no reliable estimation procedure is agreed on.

A wide area of market microstructure theory has been left out intentionally, namely the theories of market design. Comparing the effects of different trading rules and advices for the optimal design of such rules are an important issue that has attracted increased attention in recent years. Some of the questions raised in market design have been mentioned in this chapter when comparing different market forms, e.g. monopolistic and competitive market makers, and we will take up some aspects in chapter 6.

Market microstructure theories are not only important for understanding the trading process, but they also have implications for asset pricing. A large bid-ask spread gives rise to high transaction costs, which reduce the return of investors. As has been pointed out in AMIHUD AND MENDELSON (1986), we should expect the return of an asset to be increasing in the spread, a result they confirm empirically. Further we should observe a negative relationship with the average holding period of the asset, i.e. the frequency with which these costs have to be beared. If we assume trading volume to be negatively related to the average holding period, we should find a positive relationship between trading volume and the spread, which does not arise from increased adverse selection costs. BRENNAN AND SUBRAHMANYAM (1996) also point out that the reduced liquidity of the market originating from the presence

of informed investors, gives rise to higher transaction costs, hence we should find a negative relationship between the liquidity of markets and the spread, a result they also confirm empirically.

In the literature thus far in nearly all cases market participants are either assumed to act competitively or granted a monopoly. The next chapter will derive some results with imperfect competition between market participants and show how strategic behavior of market makers can affect quoted prices.

Chapter 5

Strategic behavior of market makers

The last chapter reviewed the current literature on market microstructure theory. We saw how the presence of different types of costs, inventory and adverse selection costs, affect the price setting behavior of market makers. However, the scope of this literature has been limited to market makers behaving either competitively or being granted a monopoly. As described in chapter 3, market makers on the NASDAQ are not granted a monopoly, but for most assets a relatively small number of market makers is registered, such that it is not reasonable to assume competitive behavior. A similar setting can be found on other dealer markets as well, e.g. the London Stock Exchange (LSE) or the Vienna Stock Exchange (VSE). For this reason in this chapter we will consider models allowing market makers to behave strategically and investigate the impact on the price setting behavior.

The appropriate framework for modeling strategic behavior is game theory, where the prisoner's dilemma is the most important setting in the context used here.[1] For this reason we will at first review those elements from game theory that are needed in the analysis. After introducing the basic model and deriving benchmark equilibria, the competitive and cooperative equilibria in chapter 5.2, we will in chapter 5.3 consider strategic behavior in the price setting of market makers in an economy with a single asset. After briefly describing the influence of order preferencing arrangements in chapter 5.4, we generalize our findings for an economy with multiple assets

[1] AXELROD (1984) was the first to investigate the prisoner's dilemma in detail and show under which conditions strategic behavior of individuals gives rise to noncompetitive outcomes.

in chapter 5.5 and finally consider coordination devices for market maker behavior in chapter 5.6, which also contains empirical support for the presence of strategic behavior.

Many results presented in this chapter have not previously been presented in the literature. For this reason we give detailed proofs for all of our findings, where some longer proofs are found in appendix E.

5.1 Game theoretic foundations

When considering competitive behavior of market participants, it is assumed that they can determine their payoff only through their actions and that no other market participant can affect this payoff. The same is true for monopolistic market participants. Although early models of oligopolistic markets also took into account the influence the behavior of other market participants has on payoffs, e.g. COURNOT (1838), BERTRAND (1883) or EDGEWORTH (1897), it was not before the publication of the book *Theory of Games and Economic Behavior* by John von Neumann and Oskar Morgenstern in 1944 that these aspects have been generally included in economic theories. Their work gave rise to a complete theory, called *game theory.*[2]

VON NEUMANN AND MORGENSTERN (1953, pp. 10 f.) describe a game as a situation where market participants face two kinds of variables affecting the outcome. The first variables can be controlled by each market participants himself, e.g. the quantity he produces or the prices he sets. The other variables, however, a market participant cannot control, they are the actions taken by other market participants, e.g. the quantities they produce or the prices they set. It is the existence of these latter variables that distinguishes situations to be analyzed using game theory from competitive or monopolistic settings.[3] In game theory the first group of variables

[2] See FUDENBERG AND TIROLE (1991, p. xviii).

[3] We could for completeness add a third kind of variables in both settings, variables that cannot be influenced by any market participant, e.g. moves of nature or changes in the demand schedule.

in most cases is not explicitly considered by assuming that optimal choices regarding these variables are made for each possible combination of the second group of variables, whose influences on the equilibrium are then considered.

In this section we will at first introduce the definitions and notations used in game theory before we briefly review those elements of game theory that will be needed in the remainder. It is not our aim to give a complete overview of game theory nor to address all aspects of the mentioned topics. Where needed, more detailed results will be applied later in this chapter at the appropriate place.

5.1.1 Definitions and notations[4]

As game theory has developed its own *termini technicii* which are distinct from those in other fields of economics, we have at first to introduce them. A *player* is a market participant whose outcome is affected by the behavior of other players. Market participants not affected by the behavior of other market participants or not affecting other market participants are not regarded as players. An example may illustrate this distinction: suppose a market with two suppliers of a good, i.e. a duopol in Cournot competition, and a very large number of perfectly competitive demanders for this good. The quantity offered by one supplier affects the prices and therewith also the profits made by the other supplier, hence the two suppliers are players. Demanders are also affected by the quantity offered through the equilibrium price, but due to their large number a single demander cannot affect total demand and therewith the profits of the suppliers. For this reason demanders are no players. Obviously, things change in the presence of only few demanders. In game theory we are only concerned with the behavior of players.

A certain action chosen by a player is called his *strategy*. We distinguish two kinds of strategies: pure and mixed strategies. *Pure strategies* are those strategies where

[4] This section is based on VON NEUMANN AND MORGENSTERN (1953, pp. 48 ff.) and FUDENBERG AND TIROLE (1991, pp. 4f. and pp. 70 f.) and has been adapted to the needs of this chapter.

a player chooses a single well defined action with certainty, whereas *mixed strategies* are characterized by a probability distribution over a set of pure strategies and a player chooses actions randomly according to this probability distribution.[5] In general, games are conducted in a sequence of *moves*, i.e. players choose their strategies in subsequent steps.[6] Strategies and outcomes chosen in previous moves are called the *history* of the game.

The outcome of a game is called the *payoff*, depending on economic considerations these payoffs can be utility or profits, for example. Having introduced these terms we can now more formally define the elements of a game.

As already mentioned, games are generally conducted in a sequence of moves, indexed $\nu = 1, 2, \ldots$ Which player(s) have to choose a strategy and which part of the history of the game they know in move ν is called the *rule* of the game. This rule can be allowed to depend on the history itself and belongs to a set $\mathcal{R}_\nu$. The entire set of rules is defined as $\mathcal{R} = \times_\nu \mathcal{R}_\nu$.

We further define a set of players participating in the game, $\mathcal{P}$. It is assumed that this set is finite, i.e. $\mathcal{P} = \{1, 2, \ldots, N\}$ with $2 \leq N < \infty$. As the players themselves are not important for our purpose, we can identify the set of players by the number of players, i.e. N. Each player $i \in \mathcal{P}$ has a strategy space $\mathcal{S}_\nu^i$ of pure strategies from which he can choose his strategy in move ν. We can allow this strategy space to depend on the history of the game. If a player has not to choose a strategy in move ν, we define $\mathcal{S}_\nu^i = \emptyset$. Denote $\mathcal{S}_\nu = \times_{i \in \mathcal{P}} \mathcal{S}_\nu^i$ and $\mathcal{S} = \times_\nu \mathcal{S}_\nu$. The history space is $\mathcal{H}_\nu = \times_{\tau=1}^{\nu-1} \left\{ \mathcal{S}_\tau, \times_{i=1}^{N} u_\nu^i \right\}$, where u_ν^i is the payoff of player i in move ν, which we define in an instant.

Finally we have to define a payoff function for each player $i \in \mathcal{P}$ in move ν, u_ν^i. We denote the (mostly neglected) set of variables entirely controlled by this player with

[5] Pure strategies can be interpreted as a special case of mixed strategies with the probability distribution being degenerated by having positive values only for a single action.

[6] Games with only a single move or a simultaneous move of several players are one of the most common games and will be considered in more detail below.

$\mathcal{V}_\nu^i$. Therewith the payoff function is defined as a mapping $u_\nu^i : \mathcal{V}_\nu^i \times \mathcal{H}_\nu \times \mathcal{S}_\nu^i \mapsto \mathbb{R}$.[7] At last we define the matrix of these functions to be $\mathcal{U} = (u_\nu^i)_{i\in\mathcal{P}}^{\nu=1,2,\ldots}$.

A game Γ therewith has been found to consist of four elements: a set of players $\mathcal{P}$, characterized by the number of players N, a strategy space $\mathcal{S}$, the payoff functions $\mathcal{U}$ and the rule of the game $\mathcal{R}$. Hence we can write a game as $\Gamma(N, \mathcal{S}, \mathcal{U}, \mathcal{R})$.

To complete our notations, we define the strategy space for mixed strategies by Σ_ν^i and $\Sigma_\nu = \times_{i\in\mathcal{P}}\Sigma_\nu^i$, $\Sigma = \times_\nu \Sigma_\nu$. The strategy space of all players $j \in \mathcal{P}$ with $i \neq j$ will be denoted $\mathcal{S}_\nu^{-i} = \times_{j\in\mathcal{P}, j\neq i}\mathcal{S}_\nu^j$ and $\Sigma_\nu^{-i} = \times_{j\in\mathcal{P}, j\neq i}\Sigma_\nu^j$.

For the remainder of this chapter it is not necessary to consider games as general as described above, we will restrict our attention to games in *normal form*. This reduces the rule of the game to all players making their choices in a simultaneous single move and history does not affect the outcome. The only part of the rule that may vary across our models, is the information players have on past moves in those cases where we consider repeated games, such that we can neglect the rule $\mathcal{R}$ in the remainder when having stated the information structure at the beginning. We further do not use the subindex ν to identify the move for notational simplicity.

Although a special case, but of importance for our analysis will be two-player games, i.e. games with $N = 2$. For simplifying notations we assume that those variables under complete control of a player are chosen optimal, hence we can neglect them in the payoff function. Therewith the payoff function can be written as $u^i(s_1, s_2)$ with $i \in \{1, 2\}$ and $(s_1, s_2) \in \mathcal{S}$. Furthermore the strategy space will be assumed to restrict each player to choose between the same two strategies: $\mathcal{S} = \{s^1, s^2\} \times \{s^1, s^2\}$.

The typical representation of such a game is shown in figure 5.1. Each cell corresponds to a strategy pair $(s_1, s_2) \in \mathcal{S}$. The entries denote the payoffs the individual players receive with the first entry being the payoff of player 1 and the second entry the payoff of player 2 if the corresponding strategies are chosen.

[7] It is common in game theory to neglect those set of variables that do not affect the outcome.

Player 1 \ Player 2	s^1	s^2
s^1	$u^1(s^1, s^1)$, $u^2(s^1, s^1)$	$u^1(s^1, s^2)$, $u^2(s^1, s^2)$
s^2	$u^1(s^2, s^1)$, $u^2(s^2, s^1)$	$u^1(s^2, s^2)$, $u^2(s^2, s^2)$

Figure 5.1: A two-player game in normal form

5.1.2 The concept of a Nash equilibrium

A strategy $s^i \in \mathcal{S}^i$ is said to be *strictly dominated* for a player $i \in \mathcal{P}$ if there exists a $\sigma^i \in \Sigma^i$ such that for all $s^{-i} \in \mathcal{S}^{-i}$ we find $u^i(s^i, s^{-i}) < u^i(\sigma^i, s^{-i})$. Because mixed strategies are linear combinations of pure strategies with the density or probabilities as weights in the case of a continuous and discrete strategy space, respectively, we can easily demonstrate that this inequality also holds for the players using mixed strategies, i.e. replacing $s^{-i} \in \mathcal{S}^{-i}$ by $\sigma^{-i} \in \Sigma^{-i}$ and $s^i \in \mathcal{S}^i$ by some $\sigma^i \in \Sigma^i$.[8]

We can now apply the concept of strict dominance recursively to eliminate all strategies being dominated and get those strategies that are potential equilibrium strategies. Unfortunately, this method, which is also known as *iterated strict dominance*, does in general not solve for an equilibrium, it only encompasses any equilibrium. Therefore the equilibrium cannot be defined as the result of such a procedure.

NASH (1950) introduced an equilibrium concept which has become known as the *Nash equilibrium*. A strategy profile $(\sigma^{i,*}, \sigma^{-i,*}) \in \Sigma$ is a Nash equilibrium if for all players $i \in \mathcal{P}$ and all $s^i \in \mathcal{S}^i$, we find $u^i(\sigma^{i,*}, \sigma^{-i,*}) \geq u^i(s^i, \sigma^{-i,*})$. A Nash

In the remainder we will also use this convention to simplify notations.

[8] See FUDENBERG AND TIROLE (1991, p. 7).

equilibrium therewith is a strategy profile where each player responses optimal to the other players' strategies. If $(\sigma^{i,*}, \sigma^{-i,*}) \in \mathcal{S}$ this equilibrium is called a Nash equilibrium in pure strategies, otherwise a Nash equilibrium in mixed strategies. The linearity of the payoffs in the probability function of mixed strategies allows to compare any strategy profile with another pure strategy rather than a mixed strategy without changing the argument.[9]

A Nash equilibrium in mixed strategies exists under very general conditions. NASH (1951) showed its existence if the set of pure strategies is finite. If the strategy space $\mathcal{S}^i$ is a nonempty compact set[10] of a metric space[11] and the payoff functions u^i are continuous in the strategy profiles for all $i \in \mathcal{P}$, GLICKSBERG (1952) proved the existence of a Nash equilibrium in mixed strategies. DASGUPTA AND MASKIN (1986) showed the existence of a Nash equilibrium in mixed strategies also for discontinuous payoff functions under certain regularity conditions.

These results ensuring the existence of a Nash equilibrium under very general conditions, made this concept very popular in game theory. Most contributions make use of a Nash equilibrium and its modifications. Although the existence of a Nash equilibrium in most cases is ensured, it has not to be unique nor stable to minor payoff perturbations. However, WU AND JIANG (1962) showed that almost all games with a finite pure strategy space have robust Nash equilibria. WILSON (1971) showed that almost all of these games have a finite and odd number of Nash equilibria, a result also obtained for most games with infinite pure strategy spaces.

5.1.3 Subgame perfection[12]

As we saw that Nash equilibria will in general not be unique, it is useful to distinguish between equilibria being more reasonable than others and those being less reasonable. This distinction would allow us to reduce the number of equilibria we

[9] See FUDENBERG AND TIROLE (1991, p. 11).
[10] See Appendix D.1.1 for a definition of compact sets.
[11] See Appendix D.1.3 for a definition of metric spaces.
[12] See FUDENBERG AND TIROLE (1991, pp. 92 ff.).

have to consider. With his subgame perfection SELTEN (1965) introduced a criterion to distinguish between Nash equilibria.[13]

At first we have to define a subgame.[14] Consider a game $\Gamma(N, \mathcal{S}, \mathcal{U})$ and assume that the strategy space of $1 \leq k < N$ players is reduced to a single strategy, hence we have a new strategy space $\mathcal{S}' \subsetneqq \mathcal{S}$.[15] We then can define a new game $\Gamma'(N, \mathcal{S}', \mathcal{U})$, which we call a *subgame* of the initial game Γ. This subgame can now be analyzed as any other game.[16]

Assume now that a player $j \in \mathcal{P}$ expects a certain equilibrium strategy $\sigma^{-i,*} \in \Sigma^{-i}$ to be applied by the other players, hence he chooses his equilibrium strategy $s^{j,*} \in \mathcal{S}^j$ corresponding to this equilibrium. Provided the other players know his inferences, they can take his behavior as given and set $\mathcal{S}'^j = \{s^{j,*}\}$. They will base their decisions not on the entire game Γ, but only on the subgame with $\mathcal{S}' = \times_{i=1}^{j-1} \mathcal{S}^i \times \mathcal{S}'^j \times_{i=j+1}^{N} \mathcal{S}^i$. The remaining players $i \in \mathcal{P} \backslash \{j\}$ will now choose their strategies in this subgame. If the Nash equilibrium of this subgame consists of the same strategies as the Nash equilibrium of the entire game, i.e. players do not deviate from their strategies, the equilibrium is reasonable. An equilibrium with this property is called subgame perfect.

We can formally define that a Nash equilibrium $\sigma^* \in \Sigma$ is a *subgame-perfect equilibrium* if it is also a Nash equilibrium for every subgame.

While the Nash equilibrium ensures that not a single player has incentives to deviate from the equilibrium strategy, subgame perfection ensures that even multiple players have no incentives to deviate jointly.

[13] In chapter 5.6.1 we will consider another criteria, focal points, which is based on a different reasoning.

[14] The concept of subgame perfection has been developed for games with multiple moves, so called *extensive form games*. We have adapted this concept here to fit for games in normal form.

[15] Without changing the arguments we could also assume players to be fixed to apply mixed strategies.

[16] In extensive form games a subgame is defined as a game starting in move $\nu \geq 1$ rather than in move 1, taking any previous moves given.

A major critique of this concept has been that it requires all players to agree on their inferences on the behavior of the other players, i.e. the choice of the equilibrium strategy. Another problem arises in cases where subgames have multiple Nash equilibria, here it is not possible to determine whether an equilibrium is subgame perfect. Applying subgame perfection to these equilibria does not necessarily solve this problem.

As easily can be shown, for every game with a Nash equilibrium, at least one is subgame perfect. However, in general more than one Nash equilibrium is subgame perfect. Therewith the number of equilibria to be considered can be reduced, but it is not ensured that only a single equilibrium is reasonable.

5.1.4 Bayesian equilibrium

Thus far we implicitly assumed the payoff functions of all players to be common knowledge. Now we will suppose that players know their own payoff function, but not those of the other players. We suppose for every $i \in \mathcal{P}$ that the payoff function u^i is taken from a commonly known set $\mathcal{F}^i$. Furthermore, the payoff function is determined randomly from a commonly known objective distribution function f^i over $\mathcal{F}^i$, hence it is a random variable.[17]

To simplify notations, we define a bijective mapping $\eta^i : \mathcal{F}^i \mapsto \Theta^i$ for each $i \in \mathcal{P}$ and call $\theta^i \in \Theta^i$ the type of player i. Denote further $\Theta = \times_{i \in \mathcal{P}} \Theta^i$, $\Theta^{-i} = \times_{j \in \mathcal{P}, j \neq i} \Theta^j$, $\theta \in \Theta$ and $\theta^{-i} \in \Theta^{-i}$. Therewith we can state the types of the players, and due to η^i also the payoff functions, to be drawn from the distribution function $f^i(\theta)$. Let $F^i(\theta)$ denote the cumulative distribution function and $f^i(\theta^{-i}|\theta^i)$ the conditional probability of player i's inferences of the other players' types.

HARSANYI (1968) proposes to transform this game of incomplete information into a game of imperfect information. He therefore introduces an additional player, he calls

[17] We could instead also use subjective probability functions and interpret the distribution function as the beliefs of the players without changing the argument.

nature, who determines randomly the types of the players according to f^i, hence nature chooses a given mixed strategy. In this expanded game the strategy space of each player $i \in \mathcal{P}$ changes to the set $\mathcal{S}^i_{\Theta^i}$ of mappings from Θ^i to $\mathcal{S}^i$, $\gamma_i : \times^i \mapsto \mathcal{S}^i$, where $\gamma^i \in \mathcal{S}^i_{\Theta^i}$. In other words, each player chooses the optimal function given his own type. The payoff function now also depends on the types of players, hence $u^i : \mathcal{S}^i_{\Theta^i} \times \Theta^i \mapsto \mathbb{R}$.

Define the *Bayesian equilibrium* to be the Nash equilibrium of this expanded game. A strategy $\gamma^*(\theta)$ is a Bayesian equilibrium if for all players $i \in \mathcal{P}$ and all $\gamma^i \in \mathcal{S}^i_{\Theta^i}$ we find that

$$\int_{\theta^{-i}} u^i\left(\gamma^{i,*}, \gamma^{-i,*}, \theta\right) dF^i(\theta^{-i}|\theta^i) \geq \int_{\theta^{-i}} u^i\left(\gamma^i, \gamma^{-i,*}, \theta\right) dF^i(\theta^{-i}|\theta^i). \tag{5.1}$$

As the Bayesian equilibrium is a Nash equilibrium, its existence for most cases is ensured and subgame perfection can be applied to reduce the number of equilibria to be considered. We can interpret the initial game as a subgame of the expanded game, where nature is assumed to play a mixed strategy as determined by f^i.

5.1.5 Repeated games with complete information[18]

We now consider a game with only a single simultaneous move by all players, $\Gamma(N, \mathcal{S}, \mathcal{U})$. We will call this game the *stage game* or *constituent game*. Let us then define a new game Γ', where in each of its $1 \leq T \leq \infty$ moves the stage game is played. This game is called a *repeated game*, as the stage game is repeated unchanged T times.

The strategy space of the repeated game becomes $\mathcal{S}' = \times_{t=1}^T \mathcal{S}$ and the space of mixed strategies is defined by $\Sigma' = \times_{t=1}^T \Sigma$. As the history of the game, $\mathcal{H}_t$, may influence the choice of strategies, we define the chosen strategy in move t as a mapping $\sigma_t : \mathcal{H}_t \mapsto \Sigma$ for any $t = 1, 2, \ldots, T$. Let us further denote $h_t \in \mathcal{H}_t$. The payoff function becomes the present value of future payoffs in each move. With

[18] This section is based on FUDENBERG AND TIROLE (1991, pp. 146-165).

$\rho \in [0,1]$ denoting the per period discount factor for future payoffs[19] and $\sigma \in \Sigma$ we get the payoff function as

$$g^i(\sigma) \equiv \sum_{t=0}^{T} \rho^t u^i(\sigma_t(h_t)). \tag{5.2}$$

The rule of the repeated game is, besides the knowledge of the history, characterized by the number of repetitions, T. With $\mathcal{G} = (g^i)_{i \in \mathcal{P}}$ we can therewith describe the repeated game by $\Gamma'(N, \mathcal{S}', \mathcal{G}, T)$.

It is possible to relate the Nash equilibria of the repeated game to those of the stage game. As shown in FUDENBERG AND TIROLE (1991, p. 149), we find that it is a subgame perfect Nash equilibrium of the repeated game to choose a Nash equilibrium of the stage game in every move.[20] Furthermore, if the stage game has multiple Nash equilibria, choosing any sequence of these equilibria is a subgame perfect Nash equilibrium of the repeated game.

Although we have found some Nash equilibria of the repeated game, we are interested in characterizing the entire set of Nash equilibria. Let us therefore at first consider *infinitely repeated games*, i.e. games with $T = \infty$.

It is easy to demonstrate that in every stage game there exists a payoff $\underline{u}^i$ for each player $i \in \mathcal{P}$ that a player can achieve, regardless of the choice of other players. This payoff, defined by

$$\underline{u}^i = \min_{\sigma^{-i} \in \Sigma^{-i}} \max_{\sigma^i \in \Sigma^i} u^i(\sigma^i, \sigma^{-i}), \tag{5.3}$$

is called player i's *reservation utility* or *minmax value*. As a player always can achieve his reservation utility, it can be demonstrated that any Nash equilibrium of the stage game must give him a payoff of at least $\underline{u}^i$ and in the repeated game he must at least receive this payoff in every move. A payoff vector $u = (u^1, u^2, \ldots, u^N)$

[19] The discount factor allows to calculate the present value of future payoffs. The lower the discount factor the less future payoffs are weighted compared to current payoffs, hence it is a measure for the patience of the players. For t periods ahead the appropriate discount factor to determine the present value is ρ^t.

[20] As the repeated game is an extensive form game, a subgame here we define as any game starting in move $1 \leq \tau \leq T$.

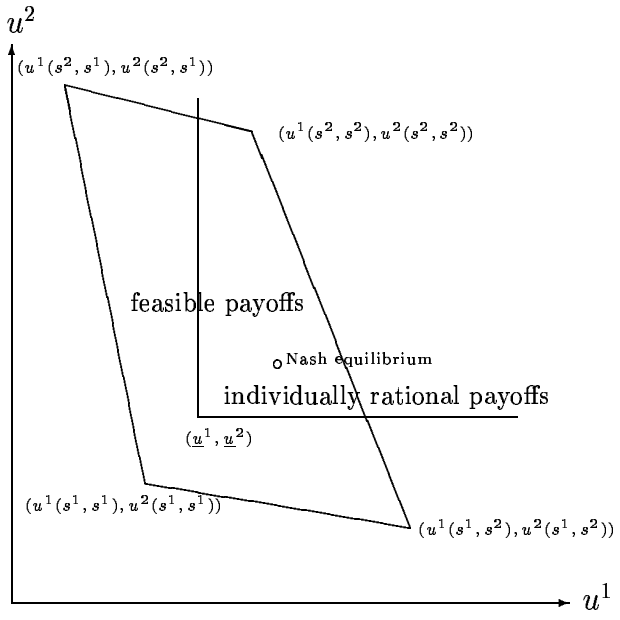


Figure 5.2: Individually rational and feasible payoffs in the stage game

with $u^i(\sigma^i,\sigma^{-i}) \geq \underline{u}^i$ for every $i \in \mathcal{P}$ is called *individually rational*. The individually rational payoffs in the infinitely repeated game are therewith all those with a payoff of $\frac{\underline{u}^i}{1-\rho^i}$ and higher.

The set of *feasible* payoffs, i.e. those payoffs that can be achieved in the repeated game are a convex set[21] under certain additional assumptions. If only pure strategies are used or otherwise the discount factor is sufficiently close to 1, the feasible payoffs are always convex. FUDENBERG AND MASKIN (1986) and FUDENBERG AND MASKIN (1990) present assumptions that ensure this convexity in more general settings. For the remainder we will assume the set of feasible payoffs to be convex by concentrating on pure strategies. Figure 5.2 shows an example for a set of individually rational and feasible payoffs in the case of a two-player game as presented in figure 5.1.

We now suppose the following behavior of players: a player $i \in \mathcal{P}$ chooses a strategy $\sigma^i \in \Sigma^i$ in period 0, such that $u^i(\sigma^i,\sigma^{-i}) = v^i$ for some $\sigma^{-i} \in \Sigma^{-i}$. He chooses σ^i in subsequent moves as long as the other players choose σ^{-i}, and hence he receives the payoff v^i. Player i could also choose a strategy $\sigma^{i'}$ such that $u^i(\sigma^{i'},\sigma^{-i}) > v^i$. The

[21] See Appendix D.1.2 for the definition of convexity.

other players are assumed to receive a lower payoff than with his previous choice. In the next move they react by playing a strategy profile $\sigma^{-i'} \in \Sigma^{-i}$ such that for any $\sigma^i \in \Sigma^i$ it is $u^i(\sigma^i, \sigma^{-i'}) < v^i$ for all remaining moves of the game, i.e. they punish the deviation of player i.

We assume here that the strategies chosen by a player can be observed *ex post* by all other players. We further suppose that the discount factor ρ and payoff functions are common knowledge.

If the players react in this way, it can be shown that for any feasible and individual rational payoff vector $v = (v^1, v^2, \ldots, v^N)$ there exists a discount factor $\rho_0 \in [0, 1[$ such that for all $\rho_0 < \rho \leq 1$ a Nash equilibrium of the repeated game with this payoff vector v exists. If $\rho \to 1$ all feasible payoffs can be achieved. This result is known as the *folk theorem* in game theory.[22]

The intuition behind the folk theorem is that the one period gain from a deviation is outweighed by the long run losses in all future moves, provided players are sufficiently patient, i.e. their discount factors for future payoffs are sufficiently high.

We may illustrate the folk theorem with the stage game shown in figure 5.3, which is also known as the *prisoner's dilemma*. The only Nash equilibrium of this stage game is (D, D). However, the strategy (C, C) would give both players a higher payoff.

Consider now the infinitely repeated prisoner's dilemma and suppose player 1 chooses to play C in move 1 expecting player 2 also to do so and hence receive a payoff of $u^1(C, C) = 1$. If player 2 plays D instead, he receives $u^2(C, D) = 2$, but in the next move player 1 will react and play D for the rest of the game. Hence to achieve his reservation utility of 0, player 2 also plays D and receives a payoff of $u^2(D, D) = 0$. According to (5.2) he receives a payoff of $2 + \frac{\rho}{1-\rho}0 = 2$ when defecting. If he had played C instead in the first move, he would have received 1 in each move, hence a payoff of $\frac{1}{1-\rho}$ for the entire game.

[22] There exists a large number of different versions of the folk theorem in the literature using various frameworks, see e.g. FRIEDMAN (1971), AUMANN AND SHAPLEY (1976), FUDENBERG AND MASKIN (1986) or FUDENBERG AND TIROLE (1991, p. 152).

Player 1 \ Player 2	C	D
C	1, 1	-1, 2
D	2, -1	0, 0

Figure 5.3: The prisoner's dilemma

The strategy to play C is better for player 2 if $\frac{1}{1-\rho} > 2$ or $\rho > \frac{1}{2}$. As player 1 makes the same considerations, the payoff $(1,1)$ in each move is achievable if $\rho > \rho_0 = \frac{1}{2}$, i.e. players are sufficiently patient.

Thus far we only assumed that a defecting player is punished by the other players through their choice of another strategy profile which gives the defecting player a lower payoff. As this in general also reduces the payoffs of those players punishing, it is important to determine whether a punishment is credible. Hence we have to find "optimal" and credible punishment strategies.

Punishment strategies are only credible if the strategies applied are reasonable, i.e. if they are subgame perfect equilibria. As mentioned above, to choose a Nash equilibrium of the stage game in every move is a subgame perfect equilibrium of the repeated game, therewith also of any subgame of the infinitely repeated game. FRIEDMAN (1971) proofs the *perfect folk theorem* or *Nash-threats folk theorem* by applying a Nash equilibrium of the stage game after a player defected.

Unfortunately, as we have seen before, in general a stage game has many Nash

equilibria, which of these equilibria has to be chosen depends on the payoff they generate for the deviating player. It is easy to see that the payoffs supported for any given discount factor $\rho < 1$ are higher, the smaller the payoffs after a defection are. The punishment strategies giving a deviator the smallest possible payoff therefore ensure support of the highest payoffs. FUDENBERG AND LEVINE (1983) show the existence of such a *worst subgame perfect equilibrium* for each player $i \in \mathcal{P}$ if the strategy space is finite. ABREU (1988) shows this existence also for strategy spaces that are compact subsets of finite dimensional Euclidian spaces[23], continuous payoff functions and stage games that have a Nash equilibrium in pure strategies.

These results suggest that sufficiently patient players can achieve a higher payoff in every move of a repeated game than in the stage game Nash equilibrium. The more patient players are, the higher these payoffs can become. The threat of switching to the worst subgame perfect Nash equilibrium for the remaining moves of the game prevents players to deviate from the strategy supporting this payoff, although it is profitable in the current period. As no binding agreements have to be used to enforce that players choose strategies giving all players a high payoff, nor the identity of a defecting player has to be revealed[24], such a behavior is called *implicit collusion* or *tacit collusion*.

In the case of *finitely repeated games*, i.e. $T < \infty$, these results in general do not hold. BENOIT AND KRISHNA (1987) show that if the payoff in the Nash equilibrium of the stage game exceeds reservation utility and $\rho = 1$, then for $T \to \infty$ all feasible and individually rational payoffs can be achieved. For stage games with at least two Nash equilibria, BENOIT AND KRISHNA (1985) proof that for sufficiently large T all feasible and individually rational payoffs can be achieved. NEYMAN (1999) shows that the folk theorem holds in cases where players do not know the number of repetitions with certainty.

The reason for this difference between finitely and infinitely repeated games is that

[23] See Appendix D.1.3 for a definition of Euclidean spaces.

[24] It is sufficient to observe that some player did deviate and choose the strategy corresponding to the worst Nash equilibrium in the remaining moves.

in finitely repeated games it is rational to deviate in the last move as no punishment can be used. Knowing this, it is rational to deviate in the previous move as it is known that in the last move all will defect, and so forth. Hence it would in general be rational to defect in all moves unless some restrictions are imposed.

5.1.6 Repeated games with imperfect public information[25]

In the previous section we assumed that the players can observe the strategies chosen by the other players. In many situations, however, it is more reasonable to assume that not the actions can be observed, but only the outcomes, e.g. market prices, but not quantities offered. This outcome is called the *public information.* As long as the public information corresponds to the strategies chosen, players can infer the chosen strategies, but now assume that the outcome is affected by a random variable $\omega \in \Omega$, whose realization is not known. With $\mathcal{Y}$ denoting the set of public information, the realization of the outcome is a mapping $y : \mathcal{S} \times \Omega \mapsto \mathcal{Y}$. Hence the outcome provides only a noisy signal of the strategies chosen.

Let $p(y|\sigma)$ with $y \in \mathcal{Y}$ and $\sigma \in \Sigma$ denote the probability of outcome y if strategy profile σ is chosen.[26] The realized payoff for player $i \in \mathcal{P}$ is $r^i(\sigma, y(\sigma, \omega))$. Therewith the expected payoffs for player i for a single move are given by

$$u^i(\sigma) = \sum_{y \in \mathcal{Y}} p(y|\sigma) r^i(\sigma, y(\sigma, \omega)). \tag{5.4}$$

The history of the game, $h_t \in \mathcal{H}_t$, is no longer the strategies chosen, but the public information of past outcomes, i.e. $\mathcal{H}_t = \times_{\tau=1}^{t-1} \mathcal{Y}$.

Green and Porter (1984) analyzed this setting by defining an equilibrium in *trigger-price strategies.* Define a trigger outcome $\widehat{y} \in \mathcal{Y}$ such that as long as $y > \widehat{y}$ players continue to cooperate because they infer from the public information that no

[25] This section is based on Fudenberg and Tirole (1991, pp. 182-187).

[26] We here assume the random variable ω to be discrete. With a continuous random variable we would have to use the conditional density instead of the conditional probability and in equation (5.4) replace the sum by an integral. However, the results to be presented here do not change with a continuous random variable.

player has deviated. However, for $y < \widehat{y}$ they switch to their punishment equilibrium, assumed to be the worst subgame perfect Nash equilibrium of the stage game as in the previous section. We assume that they apply this strategy not necessarily in all remaining moves of the game, but for $0 \leq \widehat{T} \leq \infty$ moves. To reduce notations we normalize payoffs in the punishment phase to be zero without loss of generality.

Define $\lambda(\sigma, \widehat{y}) \equiv \text{Prob}(y \geq \widehat{y}|\sigma)$ as the probability that the outcome is at least the trigger outcome if strategy profile σ is chosen. If the players follow the above rule, their payoffs can easily be determined, provided that they do not vary over time.[27]

The fundamental recurrence relation from (D.7) used in dynamic programming can be used to provide a elegant formulation of this total payoff $g^i(\sigma)$. The expected current payoff is given by $u^i(\sigma)$, which corresponds to $I(x, u, t)\Delta t$ with $\Delta t = 1$. With probability $\lambda(\sigma, \widehat{y})$ the players continue to apply the cooperative equilibrium, hence the present value of future payoffs is given by $\rho g^i(\sigma)$. With probability $1 - \lambda(\sigma, \widehat{y})$ the players apply the competitive equilibrium in the next $\widehat{T}$ periods what generates zero profits. Only afterwards the players return to the cooperative equilibrium and receive a payoff with a present value of $\rho^{\widehat{T}+1} g^i(\sigma)$. Inserting these considerations we get

$$g^i(\sigma) = u^i(\sigma) + \rho \left[\lambda(\sigma, \widehat{y}) g^i(\sigma) + (1 - \lambda(\sigma, \widehat{y})) \rho^{\widehat{T}} g^i(\sigma) \right], \tag{5.5}$$

which solves for

$$g^i(\sigma) = \frac{u^i(\sigma)}{1 - \rho\lambda(\sigma, \widehat{y}) - \rho^{\widehat{T}+1}(1 - \lambda(\sigma, \widehat{y}))}. \tag{5.6}$$

The optimal strategy σ^* has to be chosen such that a deviation is not profitable. Hence we need for all $\sigma^i \in \Sigma^i$ and all $i \in \mathcal{P}$

$$g^i(\sigma^i, \sigma^{-i,*}) \leq g^i(\sigma^*). \tag{5.7}$$

The optimal trigger price equilibrium consists of a triple $(\sigma^*, \widehat{T}^*, \widehat{y}^*)$ that maximizes (5.6) subject to (5.7).

[27] Payoffs and optimal strategies changing over time could easily be incorporated at the cost of additional notation without changing the argument.

In general we find that $\widehat{T}^* < \infty$, i.e. punishments are of finite length. If one assumes the relevant outcome to be the price, which is affected by stochastic demand, besides the quantities produced, we will find that in times of low demand, e.g. in a recession, the trigger outcome is not achieved and price wars emerge, a result confirmed empirically. For example, MacDonald (2000) reports that prices for seasonal goods decrease in times of high demand and offers implicit collusion as an explanation for his finding.

All players correctly infer from incentive constraint (5.7) that no player deviated, but the punishment phase is entered as a self-enforcing reaction to announce credibly, i.e. that defection will give rise to a punishment.

5.1.7 Reputation effects[28]

Suppose that not all characteristics of a player are perfectly known to the other players. This imperfect information could relate to his payoffs, e.g. his costs are not known, or his patience. In these cases the behavior of this player cannot be predicted with certainty.

If such a player always plays in a certain way, the other players will sooner or later expect him to do him so in the future and hence adapt their strategies. The player has developed the *reputation* to choose certain actions. We model reputation by assuming that the other players do not know the type of a player, which characterize his way of playing the game. Here the type consists of the characteristics of the player, e.g. his discount factor or payoff function. The other players form beliefs about his type, i.e. assign a probability to each of his possible types, for instance through observing his actions. The higher the probability assigned to a certain type, the higher his reputation is for being of this type.

It is common to assume that the players are of two types, where it is generally believed by all other players that with probability $1 - \varepsilon$, where $\varepsilon > 0$, the player

[28] This section follows Fudenberg and Tirole (1991, pp. 384 ff.).

is "sane", i.e. behaves rationally as predicted by standard game theory. With probability ε a player is "crazy", i.e. applies another strategy not regarded as rational.[29] KREPS ET AL. (1982) apply such a framework to analyze a finitely repeated prisoner's dilemma and show that for a sufficiently large number of moves, players cooperate in nearly all moves if "crazy" players play tit-for-tat.[30]

More formally, FUDENBERG AND MASKIN (1986) showed that any feasible and individually rational payoff can be achieved for any $\varepsilon > 0$ if the number of moves is sufficiently large. They do not have to make any assumptions about how "crazy" players choose their strategies.

This result shows that reputation effects can give rise to cooperative outcomes, even for finitely repeated games.

5.2 Benchmark equilibria

This section will present the model that is used further on to analyze strategic price setting in dealer markets. We will also derive two equilibria that do not allow for strategic behavior of market makers, but will serve as benchmarks to compare the results with strategic market makers.

5.2.1 The general model

Using the notations from game theory as introduced in chapter 5.1, we will now derive the model to be applied for investigating the price setting behavior of market makers. Although very technical, the general formulation of the model will allow us to refer to this presentation for all models to be developed as it states the basic properties of its elements. As the model derived in DUTTA AND MADHAVAN (1997)

[29] We may more reasonably interpret a player to be "sane" if he behaves in the way predicted by the other players given their inferences on his type. He is "crazy" if he applies another strategy not consistent with these inferences.

[30] Tit-for-tat is a strategy where a player chooses the strategy the other player has chosen in the previous move.

forms the basis of our model, we use mostly their assumptions modifying them slightly.

We assume a dealer market with an exogenously determined number of $N \geq 1$ market makers, hence our player set is $\mathcal{P} = \{1, 2, \ldots, N\}$ with $N < \infty$. Each market maker has the obligation to quote an irrevocable price at which he is willing to buy (bid price) and sell (ask price) a single risky asset at the beginning of every trading period. These quotes are the only variables he can directly control. For reasons to be seen later, we will only concentrate our analysis on the price setting at the ask side, hence the strategy space for each market maker $i \in \mathcal{P}$ is $\mathcal{S}^i = [0, \infty[$[31] and the entire strategy space is $\mathcal{S} = \times_{i \in \mathcal{P}} \mathcal{S}^i$.[32] We further assume that there is an infinite number of trading rounds conducted, i.e. $T = \infty$, and that these trading rounds are equally spaced. Market makers have to post their quotes in a simultaneous move at the beginning of each trading round. Future payoffs are discounted with a per period discount factor of $\rho \in [0, 1]$.

Investors are assumed to trade for exogenous reasons,[33] they all agree on the fundamental value p^* of the asset, which also is common knowledge for market makers. The demand of investors to trade with the market maker[34] depends on the quoted bid and ask prices, hence they are risk averse hedgers, and a liquidity event determining the level of demand in a trading period.

The liquidity event is an independent and identical distributed random variable taken from the set $\mathcal{L} = [0, \overline{L}]$.[35] Restricting the liquidity event to nonnegative values can be justified with the observation that forces to trade cannot be negative, in the

[31] Using this strategy space corresponds to assuming that prices can be set continuously. In chapter 5.3.1 we will also consider implications of discrete prices.

[32] We use here the common assumption that the asset has limited liability and hence negative prices are not rational. Including assets with unlimited liability does not change the argument. In this case we would define $\mathcal{S}^i = \mathbb{R}$ and $\mathcal{S} = \mathbb{R}^N$.

[33] For this reason market makers face no adverse selection costs, hence the costs of market making are not increasing with trading volume.

[34] We assume that all trades have to be conducted through market makers, i.e. investors are not allowed to submit limit orders or trade directly with each other.

[35] At the cost of additional notation we could also allow the liquidity event to follow any stochastic process without changing results significantly.

extremist case they are absent, corresponding to a level of 0. The upper bound is due to the restriction that forces to trade are first limited by the extent of resources available to investors (budget restrictions) and secondly, order handling capacities of exchanges are limited, hence the effective demand reaching the exchange is limited. We finally assume that all investors face the same liquidity event and in all remaining aspects are alike.[36] We therewith found the demand for a trade at the bid and at the ask to be a mapping for each market maker $i \in \mathcal{P}$, which we assume to be continuously differentiable and convex:

$$\begin{aligned} d_a^i &: \mathcal{S}^i \times \mathcal{L} \mapsto \mathcal{D} \\ &\quad (p_a^i, L) \mapsto d_a^i(p_a^i, L) \\ d_b^i &: \mathcal{S}^i \times \mathcal{L} \mapsto \mathcal{D} \\ &\quad (p_b^i, L) \mapsto d_b^i(p_b^i, L) \end{aligned} \tag{5.8}$$

where p_a^i and p_b^i denote the bid and ask prices quoted by market maker i, L the liquidity event and $\mathcal{D} = [0, \overline{D}]$.[37] Demands for a trade at the bid and ask are assumed to be symmetric:

$$\forall \alpha \in \mathbb{R}\ \forall L \in \mathcal{L}\ \forall i \in \mathcal{P} : d_a^i(p^* + \alpha, L) = d_b^i(p^* - \alpha, L). \tag{5.9}$$

This symmetry of demands for a trade at the bid and ask allows us to concentrate on one side of the trade, the ask side in our case, as all results derived here, apply symmetrically also to the other side. Lifting this assumption would not change the results significantly, but the bid side had to be considered separately. For notational simplicity we neglect the subscript $_a$ in the remainder.

We also make the following common assumptions on the monotonicity of the demand in the quoted price as well as the size of the liquidity event:

$$\forall L \in \mathcal{L}\ \forall i \in \mathcal{P} : \frac{\partial d^i(p^i, L)}{\partial p^i} \leq 0, \tag{5.10}$$

[36] Implicitly it is assumed that the number of investors is very large such that they behave competitively.

[37] We will show below that $\mathcal{D}$ is bounded.

$$\forall p^i \in \mathcal{S}^i \ \forall i \in \mathcal{P} : \frac{\partial d^i(p^i, L)}{\partial L} \geq 0. \tag{5.11}$$

As a normalization we set further

$$\forall p^i \in \mathcal{S}^i \ \forall i \in \mathcal{P} : d^i(p^i, 0) = 0, \tag{5.12}$$

$$\forall L \in \mathcal{L} \ \forall i \in \mathcal{P} : \lim_{p^i \to \infty} d^i(p^i, L) = 0. \tag{5.13}$$

This normalization and the monotonicity gives rise to $0 = d^i(p^i, 0) \leq d^i(p^i, L) \leq d^i(p^i, \overline{L}) < \infty$ for any $L \in \mathcal{L}$ and p^i. With the additional assumption $d^i(0, L) < \infty$ for any $L \in \mathcal{L}$, we see that there exists a $\overline{D} < \infty$ such that $d^i(p^i, L) \in \mathcal{D} = [0, \overline{D}]$ for all $p^i \in \mathcal{P}$ and all $L \in \mathcal{L}$ as stated above.

The costs of market making for any market maker $i \in \mathcal{P}$, c^i, are assumed not to depend on the size of the liquidity event and to be common knowledge for all market makers. In this specification we allow market makers to face inventory and order processing costs, but no adverse selection costs, which, as we saw in chapter 4.4, are usually correlated with demand and hence the liquidity event. This final aspect is also consistent with the assumption that the fundamental value of the asset is common knowledge and trades are exogenously induced. Let us finally in line with market microstructure theory suppose costs to be no fixed costs, i.e. they are only present if market makers conduct trades.

With the usual assumption of strict price priority, a market maker only conducts a trade if he quotes the best, i.e. lowest, price. If a subset of $\mathcal{Q} \subseteq \mathcal{P}$ market makers quote the best price, each market maker $i \in \mathcal{Q}$ receives a share of ϕ^i of the entire order flow whose value is common knowledge for all market makers,[38] where $\sum_{i \in \mathcal{Q}} \phi^i = 1$. We define a mapping

$$\lambda^i : \mathcal{S} \times \mathcal{L} \mapsto [0,1] \tag{5.14}$$
$$(p, L) \mapsto \begin{cases} 1 & \text{if} \quad \forall j \in \mathcal{P}, i \neq j : p^i < p^j \\ \phi^i & \text{if} \quad \forall j \in \mathcal{Q} : p^i = p^j, \forall j \in \mathcal{P} \backslash \mathcal{Q} : p^i < p^j \\ 0 & \text{if} \quad \exists j \in \mathcal{P} : p^i > p^j \end{cases},$$

[38] We assume here that quotes set by market makers are not ordered according to time priority, but viewed as being alike. As pointed out in NASD, Inc. (2000) no dealer market applies strict time priority, such that this assumption is realistic.

where $p \in \mathcal{S}$ denotes the vector of quoted prices by all market makers $i \in \mathcal{P}$. We can allow ϕ^i not only to depend on the cardinality of the set $\mathcal{Q}$, but also on its specific composition, i.e. which market makers jointly quote the best available prices. We suppose the relevant payoff function to be the profits of a market maker:[39]

$$\begin{aligned} \pi^i \quad : \quad & \mathcal{S} \times \mathcal{L} \mapsto \mathbb{R} \\ & (p, L) \mapsto \lambda^i(p, L)(p^i - (p^* + c^i))d^i(p^i, L). \end{aligned} \tag{5.15}$$

These profits are assumed to be strictly concave in the quoted price p^i and the liquidity event L. We finally assume the demand for the service of a specific market maker, $\lambda^i(p, L)d^i(p^i, L)$, to follow the monotone demand ratio. hence for any $p^{i,*} \leq p^i$ and $L^* \leq L$ with $p^* = (p^1, \ldots, p^{i-1}, p^{i,*}, p^{i+1}, \ldots, p^N)$ we find

$$\frac{\lambda^i(p^*, L^*)d^i(p^{i,*}, L^*)}{\lambda^i(p, L^*)d^i(p^i, L^*)} \leq \frac{\lambda^i(p^*, L)d^i(p^{i,*}, L)}{\lambda^i(p, L)d^i(p^i, L)}. \tag{5.16}$$

The present value of future profits of the game is according to (5.2) given by

$$g^i(P) = \sum_{t=0}^{\infty} \rho^t \pi^i(p, L), \tag{5.17}$$

where P denotes the matrix of all quotes over time. When we define $\mathcal{G} = (g^1, g^2, \ldots, g^N)$ we now have fully characterized the game by $\Gamma(N, \mathcal{S}, \mathcal{G}, \infty)$. The next sections will evaluate two benchmark equilibria for this model following Dutta and Madhavan (1997).

5.2.2 The competitive equilibrium

We start now analyzing the equilibria of the previously described model by assuming market makers to behave competitively. We will focus on the price setting of a single market maker as well as on the lowest available price to investors, the market price p_M, i.e. $p_M = \min\{p^1, \ldots, p^N\}$. Without loss of generality we assume throughout

[39] Although market makers are risk averse, we do not use expected utility as their objective function for analytical tractability. However, we cannot expect to gain substantial new insights into the behavior of market makers from using expected utility instead, as any risks faced by the market makers from their activities are incorporated into the costs c^i.

that market makers are indexed such that $c^1 \leq c^2 \leq \cdots \leq c^N$. Furthermore, these costs are common knowledge for all market makers.

Let us at first consider the Nash equilibrium of the stage game. Market makers will not set prices such that their profits are negative, hence we get from (5.15) for all $i \in \mathcal{P}$ the reservation price:

$$p^i \geq p^i_R \equiv p^* + c^i. \tag{5.18}$$

Suppose a market maker $j \in \mathcal{P}$ quotes a price p^j fulfilling restriction (5.18). If there exists a market maker $k \in \mathcal{P}$ such that for some $p^k < p^j$ (5.18) is fulfilled, this market maker can generate a larger profit from quoting p^k than from any price above p^j. Hence the strategy to quote $p^k < p^j$ dominates. These considerations can be made for all market makers $i \in \mathcal{P}$ and we find the market maker with the lowest costs has the lowest reservation price and hence can undercut all other market makers. The best market maker[40] will undercut the price of the second best market maker by an arbitrary small amount of $\varepsilon > 0$. The lowest price the second best market maker can set, is given by $p^2 = p^* + c^2$. Due to the continuity of all functions we can neglect this small fraction the price has to be undercut and find that

$$p_M = p^1 = p^* + c^2. \tag{5.19}$$

We therewith find that the unique Nash equilibrium of the stage game is that all market makers quote their reservation prices, except the best market maker quoting the reservation price of the second best market maker (minus a fraction) and receiving the entire order flow. This result has already been derived in chapter 4.4.2, although not using a game theoretic framework. As we assumed costs not to depend on the size of the liquidity event, the price is also unaffected by the size of the liquidity event. If several market makers are the best market makers by having equal costs, they also will quote the same prices, their reservation prices.

We have presented in chapter 5.1.5 that it is a subgame perfect Nash equilibrium of the infinitely repeated game to choose the Nash equilibrium of the stage game in

[40] It is common in the literature to denote the market maker with the lowest costs of market making as the best market maker.

every move, i.e. trading period, of the repeated game. This equilibrium we will call the *competitive equilibrium.*

We can summarize these results as follows:

Theorem 1. *The competitive equilibrium exists, is unique and has the following properties:*

- *the best market maker quotes the reservation price of the second best market maker,*
- *the other market makers quote their reservation prices, and*
- *the quotes do not depend on the size of the liquidity event.*

5.2.3 The cooperative equilibrium

Let us now assume that market makers cooperate by maximizing the present value of their future joint profits. With this cooperation we can concentrate our analysis on a single, representative market maker with costs c, hence we analyze the game $\Gamma(1, \mathcal{S}^i, \mathcal{G}, \infty)$. Furthermore, the present value of future profits is maximized if in each trading round profits are maximized, hence we have to maximize equation (5.15), where obviously for all $p \in \mathcal{S}$ and all $L \in \mathcal{L}$: $\lambda^i(p, L) = 1$.

The concavity of the profits in the quoted price ensures the existence and uniqueness of such an equilibrium. We will denote the equilibrium price p^E to be a mapping

$$\begin{aligned} p^E \quad : \quad & \mathcal{L} \mapsto \mathcal{S}^i \\ & L \mapsto p^E(L). \end{aligned} \tag{5.20}$$

Suppose that for any $L, L' \in \mathcal{L}$ with $L < L'$ we find $p^E(L') < p^E(L)$, i.e. the optimal price is decreasing in the size of the liquidity event. Due to profit maximization we find that

$$\begin{aligned} (p^E(L') - p^* - c)d(p^E(L'), L') &> (p^E(L) - p^* - c)d(p^E(L), L'), \\ (p^E(L) - p^* - c)d(p^E(L), L) &> (p^E(L') - p^* - c)d(p^E(L'), L). \end{aligned} \tag{5.21}$$

Combining these two inequalities yields

$$\frac{d(p^E(L), L)}{d(p^E(L'), L)} > \frac{d(p^E(L), L')}{d(p^E(L'), L')}, \tag{5.22}$$

which with $\lambda^i(p, L) = 1$ violates the monotone demand ratio in (5.16). Therefore the optimal price has to be weakly increasing in the size of the liquidity event.[41]

As demand is monotonically increasing in the size of the liquidity event for any given price as presented in (5.11), the profits are also increasing monotonically. Because it is optimal to quote a higher price with a larger liquidity event, profits increase further from this increase in the optimal price. Hence profits have to increase with the size of the liquidity event.

We can summarize our findings on the cooperative equilibrium as follows:

Theorem 2. *The cooperative equilibrium exists and is unique with the following properties:*

- *the optimal price is increasing in the liquidity event, and*
- *profits are increasing in the liquidity event.*

We therewith found that in the cooperative equilibrium, prices are increasing in the liquidity event, while they were unaffected in the competitive equilibrium. In the next section we will now derive properties of the optimal prices under implicit collusion.

5.3 Market making with a single asset

This section will analyze market makers in strategic interaction using the framework developed in chapter 5.2.1 and contrast the results with the benchmark equilibria

[41] In chapter 4.3.3 we investigated the price setting behavior of a monopolistic market maker facing inventory costs. We also found that the fee he charges to investors, hence the quoted price, increases in the level of demand. Therefore the result derived here is consistent with the results from market microstructure theory.

presented before. At first we will derive the results from the Dutta-Madhavan model of implicit collusion. The remaining sections then will consider several generalizations of this model not previously reported in the literature.

5.3.1 The Dutta-Madhavan model of implicit collusion

DUTTA AND MADHAVAN (1997) point out that the cooperative equilibrium as described in chapter 5.2.3 is difficult to sustain, because it requires severe penalties to be applied to defectors, e.g. by refusing interdealer trading or other coordinated actions imposing high costs also on the punishing market makers. Instead they suggest to focus on incentive compatible equilibria which do not require cooperation between market makers but result from noncooperative behavior. hence they focus on noncompetitive Nash equilibria of the repeated game.[42]

In their model, DUTTA AND MADHAVAN (1997) assume that all market makers have the same costs of market making, i.e. $\forall i \in \mathcal{P} : c^i = c$. Without loss of generality they further assume $c = 0$, i.e. the absence of any costs of market making. The relevant payoffs in this repeated game are determined according to (5.2), which we will write here in a different way by using the concept of dynamic programming:[43]

$$J^i(L) = \max_{p^i}\{\pi^i(p^i, L) + \rho^i E[J^i(L)]\}, \tag{5.23}$$

where we assume the size of the liquidity event to be a random variable which is known for the current period and the distribution for the remaining periods is common knowledge. ρ^i denotes the discount factor for market maker $i \in \mathcal{P}$, which is known to all market makers. The aim of all market makers is to maximize (5.23) by choosing the optimal price p_c^i. As all market makers know the size of the liquidity event and face the same costs of market making, they will all behave alike and quote the same price in equilibrium, p_c. The only differences between market makers are

[42] The game theoretic framework of their analysis has first been analyzed in more detail in DUTTA (1995a) and DUTTA (1995b).

[43] See Appendix D.5 for a short introduction of this concept. Here the current payoffs are denoted by $\pi^i(p^i, L)$, where $u(t)$ corresponds to p^i, x to L and $\Delta t = 1$. As time is no own variable in this model, it is neglected in the notation.

the share of the order flow they receive, ϕ^i, and the discount factor, ρ^i. If these differences would induce market makers to quote different prices above competitive levels, those excluded from the order flow as a result of quoting not the best price would make zero profits in the current period. As the reservation price was shown in (5.18) not to depend on these characteristics, they could quote a lower price such that they make positive profits. Therewith the only equilibrium is that all market makers quote the same price. As long as market makers do not behave competitively we find that $\lambda^i(p_c, L) = \phi^i$ and $\mathcal{Q} = \mathcal{P}$.

The maximization of equation (5.23) has to be constrained such that defection from this pricing rule is not profitable. The current profits from defection are given by $(p_c - p^*)d(p_c, L)$. A defector would undercut the optimal price by a fraction and receive the entire order flow, i.e. $\lambda^i(p_c - \varepsilon, L) = 1$. Due to the continuity of the demand function and therewith the profits, we can reasonably neglect this fraction the price has to be undercut. The profits in future periods are zero, because, as pointed out in chapter 5.1.5, the worst subgame perfect equilibrium is chosen afterwards to punish the defector. In our case the competitive equilibrium in every trading round is this worst equilibrium, which due to the equal costs of all market makers generates zero profits as shown in theorem 1. For market makers not defecting, the expected profits are given by $\phi^i(p_c - p^*)d(p_c, L) + \rho^i E[J^i(L)]$ according to (5.23). Hence we have the following incentive constraint that has to be fulfilled while maximizing (5.23):

$$\phi^i(p_c - p^*)d(p_c, L) + \rho^i E[J^i(L)] \geq (p_c - p^*)d(p_c, L), \tag{5.24}$$

which transforms into

$$(p_c - p^*)d(p_c, L) \leq \frac{1}{1 - \phi^i}\rho^i E[J^i(L)]. \tag{5.25}$$

As the left-hand side of (5.25) equals the profits of the cooperative equilibrium if p_c denotes the unrestricted solution to the maximization of (5.23), we can use theorem 2 to state that it is increasing in the size of the liquidity event. With the right-hand side being a constant as the liquidity event is independent and identical distributed

over time, we can define a $L_c^i \in \mathcal{L}$ such that for $L = L_c^i$ relation (5.25) is fulfilled with equality. If for $L = 0$ (5.25) is violated, we define $L_c^i = 0$ and if it is fulfilled for $L = \overline{L}$, set $L_c^i = \overline{L}$.

As long as $L < L_c^i$, constraint (5.25) is not binding and the optimal price can be determined according to the maximization of (5.23), hence the optimal price p_c equals the cooperative price and therewith is increasing in the size of the liquidity event L. For $L > L_c^i$ this constraint becomes binding and in order to reduce the left hand-side, the optimal price has to decrease, reducing profits.[44] The larger L, the more p_c has to be reduced, but it always will be above the competitive price. We therewith find that for $L > L_c^i$ the optimal price is decreasing in the size of the liquidity event and below cooperative, but above competitive levels.

These results can now be inserted into (5.23) and we get with (5.25)

$$J^i(L) \begin{cases} \leq \frac{1}{1-\phi^i}\rho^i E[J^i(L)] & \text{for} \quad L \leq L_c^i \\ = \frac{1}{1-\phi^i}\rho^i E[J^i(L)] & \text{for} \quad L > L_c^i \end{cases}, \tag{5.26}$$

hence we find

$$J^i(L) \leq \frac{1}{1-\phi^i}\rho^i E[J^i(L)]. \tag{5.27}$$

We can now take expectations of L on both sides. With $E[J^i(L)] > 0$ this can be transformed into

$$\rho^i \geq \rho_0^i = 1 - \phi^i. \tag{5.28}$$

If $E[J^i(L)] = 0$, (5.27) is always fulfilled, hence the competitive equilibrium is also an equilibrium of this game. But we can easily see that with the above pricing rule for $L \leq L_c^i$ profits are positive and for $L > L_c$ cannot become negative as the competitive equilibrium would dominate such an outcome, hence we find $E[J^i(L)] > 0$.

With $\rho^i < \rho_0^i$ for any $i \in \mathcal{P}$ the only solution to (5.27) would be $E[J^i(L)] = 0$, i.e. competitive price setting. We see from (5.28) that the smaller the share of the order

[44] We showed in chapter 5.2.3 that profits are increasing with an increase of the cooperative price such that upon a decrease of the price, profits are also decreasing.

flow is, the more patient the market maker has to be to achieve noncompetitive prices. This is because the profits from a defection are relatively large compared to the profits from implicit collusion due to his small share of the order flow in collusion.

As a final property we see from (5.25) that the right-hand side increases in ρ^i, hence the more patient market makers, the larger L_c^i becomes and for $L_c^i < L$ the price increases with the patience of the market makers, a result similar to the folk theorem. As the liquidity event is bounded, it is obvious from the continuity of demand that the profits of the current period are limited to $\pi^i(p_c(\overline{L}), \overline{L}) < \infty$. We therewith find that

$$J^i(L) \leq \frac{1}{1-\rho^i}\pi^i(p_c(\overline{L}), \overline{L}) < \infty. \tag{5.29}$$

We can take expectations and insert the result into (5.25). As with $\rho^i \to 1$ in (5.29) the right hand side converges to infinity, the right hand side of (5.25) will go to infinity, making the constraint nonbinding for all L as $(p_c - p^*)d(p_c, L) = \frac{1}{\phi^i}\pi^i(p_c, L) < \infty$. Hence for $\rho^i \to 1$ we find $L_c^i \to \overline{L}$ and cooperative pricing is applied for all liquidity events. A result previously derived as the folk theorem.[45]

To determine the equilibrium, all these relations have to be fulfilled for all market makers $i \in \mathcal{P}$ and L_c is determined according to the least patient market maker to avoid any incentives for him to defect, i.e. $L_c = \min\{L_c^1, L_c^2, \ldots, L_c^N\}$. We can summarize our results as follows:

Theorem 3. *If all market makers are sufficiently patient, i.e.* $\forall i \in \mathcal{P} : \rho^i > \rho_0^i = 1 - \phi^i$ *there exists a* L_c *such that*

- *if* $L \leq L_c$ *the optimal price equals the cooperative price and increases in* L,
- *if* $L > L_c$ *the optimal price is below cooperative and above competitive levels and decreases in* L, *and*

[45] We will not consider strategies, where payoffs are traded over time between investors as the result of different time preferences. LEHRER AND PAUZNER (1999) provide a game theoretic model including such strategies.

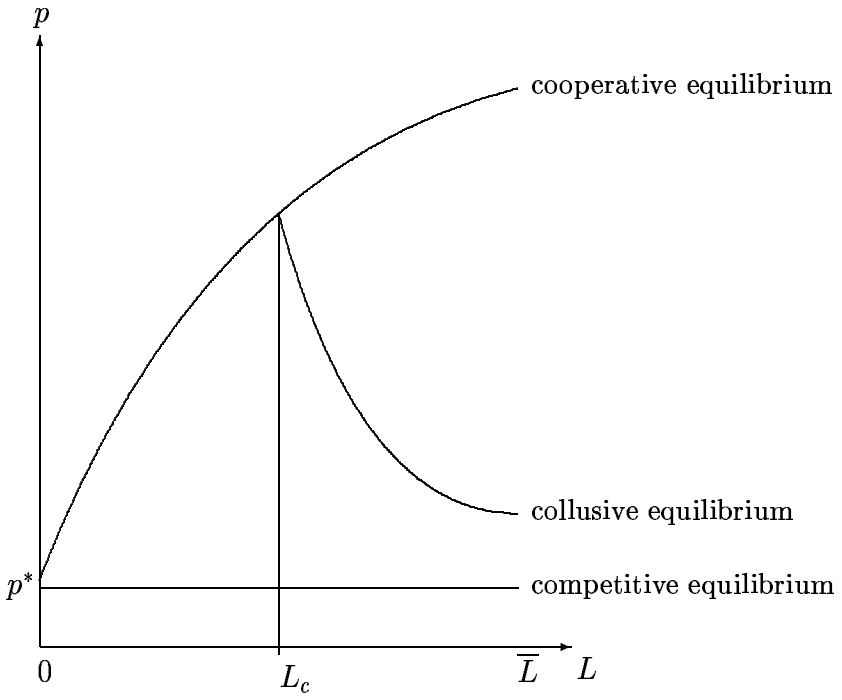


Figure 5.4: Comparison of the competitive, collusive and cooperative equilibrium

- L_c *is increasing in* $\rho = \min\{\rho^1, \rho^2, \ldots, \rho^N\}$ *for given shares of the order flow* $\phi^1, \phi^2, \ldots, \phi^N$.

If $\exists i \in \mathcal{P} : \rho^i < \rho_0^i = 1 - \phi^i$ *the only equilibrium is to set prices competitively.*

It has to be noticed that any price between the competitive price and p_c is a subgame perfect Nash equilibrium of the repeated game, hence the equilibrium is not unique, with continuously set prices there is a continuum of equilibria for each liquidity event. However, here we only consider the price giving market makers the highest profits, the optimal price.

With this pricing rule market makers quote noncompetitive prices, provided all are sufficiently patient, earning excess profits. This is achieved without cooperating in the price setting, but acting on pure self interest. We will call this behavior using the optimal prices the *collusive equilibrium*. Figure 5.4 compares the different equilibria derived thus far.

A characteristic of the collusive equilibrium is the decreasing price for large sizes of the liquidity event, while for the cooperative equilibrium it is increasing regardless of the liquidity event and does not change in the competitive equilibrium. In an empirical investigation it is more useful to consider the spread rather than the prices themselves as they vary over time with the fundamental value or market makers adjusting their quotes to inventory positions. We can use the symmetry of the results for the bid price and get the same properties as described above also for the spread.

We will now give the intuition for two further results formally shown in DUTTA AND MADHAVAN (1997) as their propositions 3 and 7. The optimal price p_c as well as L_c are not affected by the market size, i.e. the absolute scale of the demand as the relations between profits do not change, only their scales, but by the frequency of trades. With a given time preference ρ^i for a certain calendar unit, increasing the number of trades within this unit changes the discount factor between trading periods to $\rho_n^i = (\rho^i)^{\frac{1}{n}}$ with n denoting the number of trades in each calendar unit. As we can easily see, for $n \to \infty$ we find $\rho_n^i \to 1$. Hence with the results above, L_c increases and therewith prices for $L > L_c$. Furthermore, condition (5.28) is fulfilled for all market makers, even with small shares of the order flow. Therefore prices increase with trading frequency and collusion is facilitated. For this reason, we can expect to find implicit collusion even for frequently traded assets that have a large number of market makers, i.e. small ϕ^i for each market maker.

When we consider the more realistic case of discrete instead of continuous prices, i.e. prices have to be quoted on certain discrete ticks, it is no longer possible to undercut the price of a competitor only by a fraction, but by an entire tick, what reduces the profits made from defection. With a tick size of δ we have to rewrite constraint (5.36) as

$$\phi^i(p_c - p^*)d(p_c, L) + \rho^i E[J^i(L)] \geq (p_c - \delta - p^*)d(p_c - \delta, L). \tag{5.30}$$

If we approximate $\zeta(p_c - \delta, L) \equiv (p_c - \delta - p^*)d(p_c - \delta, L)$ by a first order Taylor

series around (p_c, L) we get

$$\zeta(p_c - \delta, L) = \zeta(p_c, L) - \frac{\partial \zeta(p_c, L)}{\partial p_c}\delta. \tag{5.31}$$

Defining $D = \frac{\zeta(p_c,L)}{\partial p_c}\delta \geq 0$[46] constraint (5.30) can be transformed into

$$\frac{\rho^i E[J^i(L)]}{1-\phi^i} + \frac{1}{1-\phi^i}D \geq (p_c - p^*)d(p_c, L). \tag{5.32}$$

The larger the tick size, the larger the left side of this constraint compared to continuous pricing.[47] So the profits, and therewith the collusive prices can be higher the larger the tick size is. Furthermore, we can directly see that L_c also increases. Holding L constant for $L > L_c$, the collusive prices can increase, still fulfilling (5.32). In the case of discrete prices, higher prices are sustainable than with continuous prices.

As before we can also easily show that collusion between market makers is facilitated as they can become less patient the larger the tick size is. To see this we follow the same steps as in deriving equation (5.28) and receive

$$\rho^i \geq \rho_0 = 1 - \phi^i - \frac{D}{E[J^i(L)]} \tag{5.33}$$

as the condition for the existence of a collusive equilibrium.[48] Obviously, the larger D and therewith the tick size is, the less patient market makers can become and collusion is facilitated.

In the remainder of this section we will now relax the restrictive assumptions underlying the model presented here. In chapter 5.3.2 different costs of market making will be introduced, in 5.3.3 the size of the liquidity event will not be perfectly known to market makers and in 5.3.4 they do not have perfect knowledge about strategies chosen by the other market makers.

[46] As $\zeta(p_c, L)$ denotes the profit function under cooperation, we know from theorem 2 that $\frac{\zeta(p_c,L)}{\partial p_c} \geq 0$.

[47] For continuous prices, i.e. $\delta = 0$, this constraint is equal to (5.25).

[48] In the same manner as above it is obvious that $E[J^i(L)] > 0$.

5.3.2 Different costs of market making

Assuming equal costs of market making for all market makers is very restrictive. We know from chapter 4.3 that the costs of market making also depend on inventory costs, which are determined by the inventory position held by each market maker. After every trade conducted by a market maker, this position and hence his costs change. For this reason it is more appropriate to assume market makers to have different and time varying costs of market making. In this section we will generalize the Dutta-Madhavan model to incorporate different costs of market making and consider the empirical implications of changing costs over time.

As noticed in chapter 5.2.1, with different costs of market making the best market maker is able to make a profit by quoting the reservation price of the second best market maker, even if he behaves competitively. With changing costs over time it is likely that every market maker sometimes has the lowest costs and can make these profits, hence we denote the expected profits from competitive behavior for market maker $i \in \mathcal{P}$ by $J_c^i(L)$ and define the difference in expected profits from collusive and competitive behavior by

$$K^i(L) \equiv J^i(L) - J_c^i(L). \tag{5.34}$$

As before, each market maker maximizes (5.23) subject to an incentive constraint preventing defection:

$$\phi^i(p_c^i - p^* - c^i)d^i(p_c^i, L) + \rho^i E[J^i(L)] \geq (p_c^i - p^* - c^i)d^i(p_c^i, L) + \rho^i E[J_c^i(L)], \tag{5.35}$$

where p_c^i denotes the price set by market maker $i \in \mathcal{P}$. We can transform (5.35) into the equivalent of (5.25):

$$(p_c^i - p^* - c^i)d(p_c^i, L) \leq \frac{1}{1-\phi^i}\rho^i E[K^i(L)]. \tag{5.36}$$

Using the same argumentation as in the Dutta-Madhavan model, we see that the left hand side corresponds to the expected profits in the cooperative equilibrium and

therewith is increasing in L, while the right hand side is a constant.[49] therewith for every market maker $i \in \mathcal{P}$ there exists a L_c^i such that for $L = L_c^i$ the above relation is fulfilled with equality and for $L < L_c^i$ it is not binding. As before define $L_c^i = 0$ if the constraint is violated for $L = 0$ and $L_c^i = \overline{L}$ if it is still fulfilled for $L = \overline{L}$.

Let us at first consider the case $L < L_c^i$. In this case we can conduct the maximization of (5.23) as before by maximizing expected profits in each period. The first order condition for a profit maximum becomes

$$\frac{\partial \pi^i(p^i, L)}{\partial p^i} = \phi^i \left[d^i(p^i, L) + (p^i - p^* - c^i) \frac{\partial d^i(p^i, L)}{\partial p^i} \right] = 0. \tag{5.37}$$

The concavity of the profit function ensures the second order condition for a profit maximum to be fulfilled. At first we see that the optimal price does not depend on the share of the order flow, whereas it depends on the costs of market making. The higher the costs of market making, the higher the optimal price, such that the best market maker has the lowest optimal price.

To see this suppose that the costs increase. In this case we can immediately deduct from (5.37) that for unchanged prices $\frac{\partial \pi^i(p^i, L)}{\partial p^i} > 0$, hence with the concavity of the profit function we can deduct that the optimal price has to increase.

If market makers face different costs of market making, their optimal prices as determined by (5.37) will differ, hence by quoting this price only the market maker with the lowest quote, i.e. the best market maker, would receive incoming orders. The best market maker would not set a price above his optimal price to include other market makers as this would reduce his expected profits. However, the other market makers facing higher costs of market making have an incentive to reduce their quotes to the price set by the best market maker.[50] In this case they receive

[49] The presence of costs of market making does not change this argument. By inserting $p^{**} = p^* + c^i$ for p^* in the Dutta-Madhavan model, we immediately see that the argument does not change.

[50] If the quote of the best market maker is below the reservation prices of some market makers, they would only quote their reservation prices and therewith still be excluded from the order flow. The share of the order flow for the remaining market makers increases, but as we have seen above, the optimal prices are not affected by the share of the order flow and the argumentation does not change.

a share of the order flow and can make a profit. It is not optimal for them to set a quote below the optimal price of the best market maker because he could easily use the same price, what gives rise to smaller expected profits for all market makers. Therewith it is an equilibrium for all market makers to apply the same quote as the optimal price of the best market maker and no market maker has an incentive to deviate as long as constraint (5.36) is fulfilled.[51]

Define now $L_c = \min\{L_c^1, L_c^2, \ldots, L_c^N\}$ as the liquidity event where (5.36) becomes binding for some market maker. We then see immediately that for $L < L_c$ the constraint is not binding for any market maker, hence the quoted price is determined according to (5.23) and therewith equals the cooperative price of the best market maker.[52] We know from chapter 5.3.1 that the cooperative price is increasing in the liquidity event, hence for $L < L_c$ the optimal price is also increasing in the liquidity event. This result is identical to that in the Dutta-Madhavan model with equal costs of market making.

However, for $L \geq L_c$ things slightly change. In this case prices are set such that (5.36) is fulfilled with equality for the market maker with the most binding constraint. As the right hand side also depends on the share of the order flow, it will be crucial how many market makers share the order flow. The higher the share of the order flow, i.e. the less market makers share the order flow, the less binding the constraint becomes.

If the price determined according to (5.36) is below the reservation prices of some market makers, they are excluded from the order flow as they are not able to quote the best price. For all market makers with lower reservation prices the share of the order flow is weakly increasing, hence higher prices could be applied without violating the constraint. However, the price applied by the remaining market makers has always to be below the reservation price of the excluded market maker. If they

[51] Below we also consider strategies to quote lower prices with the aim to exclude market makers from the order flow.

[52] For simplicity we assume that $p_R^N < p_c(L_c)$. Otherwise the number of market makers to be considered is reduced and the share of the order flow increases without changing the argument.

quote a higher price, he would also receive a share of the order flow, what in turn reduces the share of the order flow for the other market makers such that (5.36) is violated. Hence prices always have to be below the reservation price of the last market maker excluded.

Suppose that we face a liquidity event L_0^M such that $p_c(L_0^M, \phi^i) = p_R^{M+1}$.[53] If the liquidity event increases marginally, we see that due to (5.36) the price must fall below p_R^{M+1} and market maker $M+1$ is excluded from the order flow. For the remaining M market makers the share of the order flow weakly increases to ϕ_M^i, replacing ϕ^i in (5.36), where obviously $\phi_M^i \leq \phi_{M-1}^i$ and $\phi^i = \phi_N^i$. The price therewith can increase without violating (5.36), but with the foresaid it cannot exceed p_R^{M+1}. We find that for liquidity events not too much larger than L_0^M the optimal price is to quote p_R^{M+1} (minus a fraction) and constraint (5.36) is not binding. With the liquidity event increasing beyond L_0^M, the left hand side of (5.36) also increases, hence the maximum price fulfilling this constraint will be decreasing and there will exist a liquidity event L^M such that $p_c(L^M, \phi_M^i) = p_R^{M+1}$.[54] For $L > L^M$ constraint (5.36) becomes binding again and the price is decreasing in the liquidity event, while it is constant for $L_0^M < L \leq L^M$.[55]

These considerations give rise to the following theorem on the behavior of the collusive prices if market makers face different costs of market making:

Theorem 4. *If $\rho^i > \rho_0^i = 1 - \phi_M^i$ for all $i \in \mathcal{P}$, a unique collusive equilibrium exists with the following properties:*

- *For $L \leq L_c$ the optimal price is the cooperative price of the best market maker and is increasing in the liquidity event.*

[53] If we find that $p_c(\overline{L}, \phi^i) > p_R^{M+1}$ define $L_0^M = \overline{L}$.
[54] If we find that $p_c(\overline{L}, \phi_M^i) > p_R^{M+1}$ define $L^M = \overline{L}$.
[55] L^M is determined by the market maker for which constraint (5.36) is most binding. If the share of the order flow received by market makers excluded from the order flow is not equally distributed between the remaining market makers, this market maker can change after every excluded market maker.

- *For $L > L_c$ the optimal price is below the cooperative and above the competitive price. It equals p_R^{M+1} for $L_0^M < L \leq L^M$ and is decreasing in the liquidity event for $L^M > L \geq L_0^{M-1}$.*

If $\rho^i < \rho_0^i$ for some $i \in \mathcal{P}$, the only equilibrium is to quote competitive prices.

The formal proof on the patience of market makers for the existence of a collusive equilibrium follows the same steps as for theorem 3, only replacing ϕ^i by ϕ_M^i. It has to be noticed that if a collusive equilibrium exists with equal costs of market making, it also exists in the case of different costs of market making. This can immediately be seen as from $\phi_M^i \leq \phi^i$ we find that if $\rho^i > 1 - \phi^i$ it is $\rho^i > 1 - \phi_M^i$.

Figure 5.5 illustrates the findings of theorem 4 with $N = 5$ market makers. The bold line represents the optimal prices with different costs of market making, whereas the thin line represents the optimal prices with equal costs of market making resulting from a reservation price of p_R^1. As is obvious from (5.36), the critical liquidity event L_c is the same in the two cases if we assume all other characteristics of the market makers, i.e. patience and share of the order flow, to be the same in the two settings.

The general result from the Dutta-Madhavan model that prices are increasing in the liquidity event for $L \leq L_c$ and decreasing for $L > L_c$, remains valid also with different costs of market making. The negative relationship for $L > L_c$ becomes less pronounced due to the steps at the reservation prices of excluded market makers. Nevertheless, we can expect to find empirical support of implicit collusion by using this property of prices.

The presence of different costs of market making, however, gives market makers additional possibilities to set prices. If market makers are excluded from the order flow by quoting prices below their reservation prices, we saw above that the share of the order flow received by the remaining market makers is weakly increasing and hence expected profits are weakly increasing. Thus far we only considered market

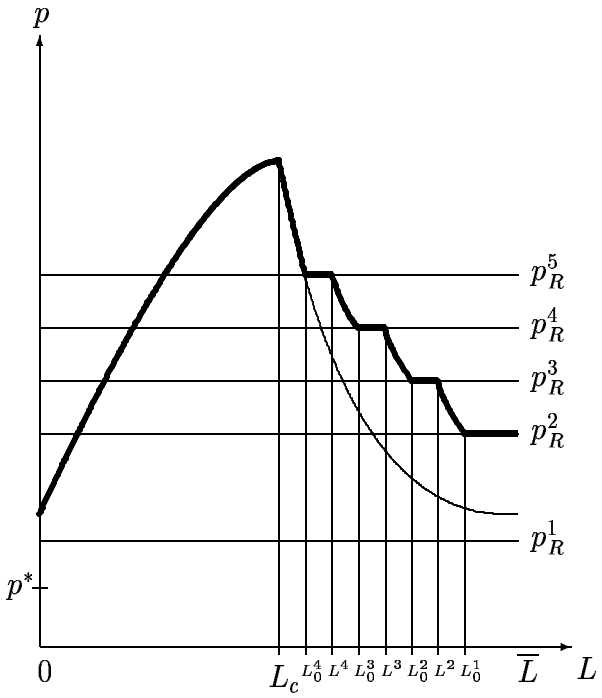


Figure 5.5: The collusive equilibrium with different costs of market making

makers to be excluded in this way because the constraint to avoid a defection would have been violated by quoting higher prices to include them in the order flow.

We can now also consider the following strategy: market makers facing low costs of market making could quote a lower price than those implied by maximizing (5.23) subject to (5.36). To quote a higher price is no equilibrium because constraint (5.36) is violated. By reducing the price, they can exclude market makers facing high costs of market making and therewith increase their share of the order flow. This increased share of the order flow will compensate them for the smaller profits made from each trade due to the lower price they set.

In this view the share of the order flow becomes a function of the quoted price, $\phi^i(p^i)$. This function will experience a discrete downward jump at every reservation price of a market maker, hence it will only be optimal to quote prices either equal to the reservation price (minus a fraction) of the last market maker excluded from the order flow or equal to the equilibrium prices according to theorem 4 if no market maker is excluded voluntary.

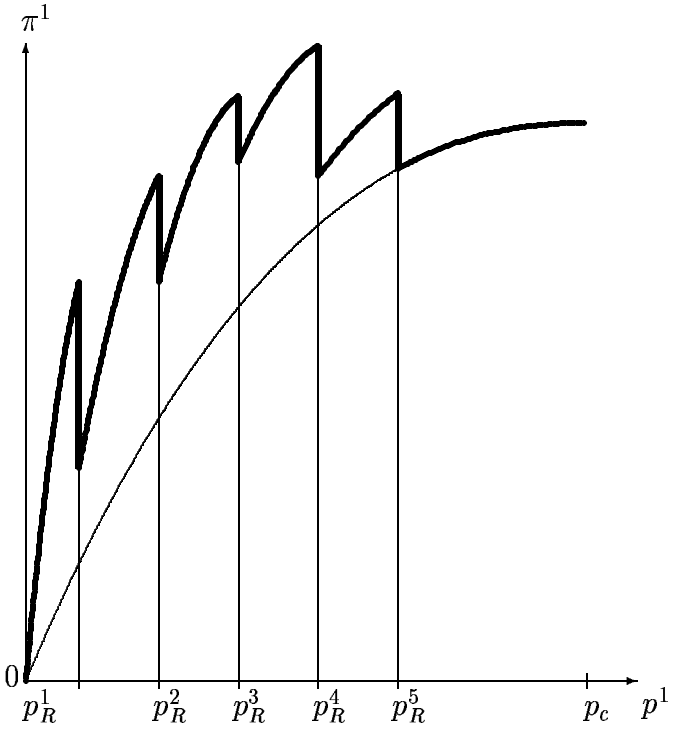


Figure 5.6: Expected profits for the best market maker with different costs of market making

In figure 5.6 the bold line shows the expected profits a market maker can achieve through different prices below the collusive equilibrium p_c, while the thin line shows the expected profits with equal costs of market making. It is obvious after the foresaid that upon excluding another market maker from the order flow, expected profits increase by a discrete amount. Due to the concavity of the profit function it also is obvious that the profit function between these steps is increasing. The prices with the highest expected profits are therefore not necessarily the collusive prices, hence the optimal price for a market maker may be below the collusive price. In the example of figure 5.6 the optimal price is $p_R^4 \ll p_c$.[56]

Applying this strategy cannot be viewed as a defective behavior by those market makers being excluded voluntarily from the order flow. First of all, it is a profit maximum if we insert the function $\phi^i(p^i)$ into (5.23), replacing the constant ϕ^i.

[56] As every market maker makes these considerations, the optimal price will always be determined by the market maker with the lowest optimal prices. This price will be applied by all market makers not excluded from the order flow.

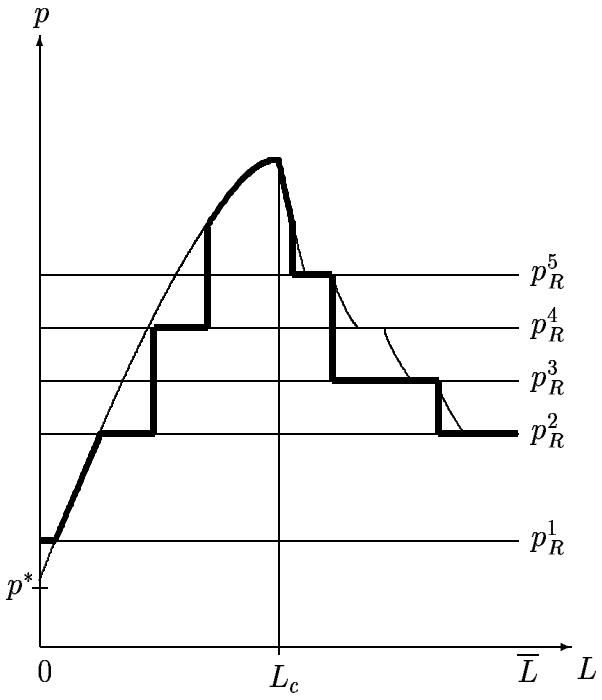


Figure 5.7: Optimal prices

Furthermore, as over time costs are changing, market makers excluded in the current period will be able to participate in future periods and therewith receive a higher expected return in these periods. Finally, in the current period they cannot react by quoting a lower price as they face costs of market making that give rise to a reservation price above the price quoted.

Figure 5.7 shows an example how the optimal prices are chosen. The bold line represents the optimal prices, while the thin line represents the prices of the collusive equilibrium. We observe that the optimal price is still increasing in the liquidity event for $L \leq L_c$ and decreasing afterwards, but as also can be seen, there will be wide areas of constant prices to be applied. In an empirical investigation we can expect to find a much less pronounced negative relationship between the price and the liquidity event for $L > L_c$.

The possibility to exclude market makers from the order flow by applying lower prices can also facilitate collusion. If there are some market makers with $\rho^i \leq \rho_0^i = 1 - \phi_M^i$, it is still possible to maintain collusive prices if we can either exclude them

		Scheme a	Scheme b	Scheme c	Scheme d
Different costs	p_R^1	100	100.05	100.075	100.1
	p_R^2	100.055	100.1	100.1	100.15
	p_R^3	100.1	100.2	100.25	100.2
	p_R^4	100.2	100.25	100.3	100.3
	p_R^5	100.25	100.3	100.4	100.4
Equal costs		100	100.05	100.075	100.1

Table 5.1: Reservation price schemes used for the simulation

from the order flow through lower prices, for which they have to face high costs of market making, or we could exclude other market makers from the order flow and therewith increase their share of the order flow such that they become willing to collude, i.e. $\rho^i > \rho_0^i = 1 - \phi_{M'}^i$.

Furthermore, as we can see from figure 5.6, the optimal price depends on the reservation prices of all market makers. Only minor modifications may change the shape of the optimal prices significantly. We will see later on in figure 5.8 an example of how much this shape can change. Therefore any shape observed in a certain period will not be stable over time as costs vary.

An empirical investigation has to use time series data and therefore will for every observation find different reservation price schemes because the costs of market making change after every transaction. The characteristics of implicit collusion is the negative relationship between the liquidity event and the quoted price for all $L > L_c$. It will therefore be useful to evaluate how this relationship is affected by different reservation price schemes. As analytical results cannot be conducted unless we make very strong assumptions on all relevant functions, we will use a simulation to illustrate the properties of the quoted price with costs of market making changing over time.

Suppose a market with $N = 5$ market makers and four possible reservation price schemes, which are shown in table 5.1. All relevant functions used have the same shape as in figures 5.4, 5.5, and 5.6.

As implicit collusion can be distinguished from competitive and cooperative price setting best by the negative relationship between the liquidity event and the quoted price for $L > L_c$, we only investigate the quoted prices for $2.5 = L_c \leq L \leq \overline{L} = 5$. Besides continuous pricing we also investigate the more realistic case of discrete prices and compared our results for different costs of market making with those for equal costs as in the Dutta-Madhavan model.

The optimal prices for a tick size of USD $^1/_8$ are shown in figure 5.8. The bold line represents the optimal prices with different costs of market making while the thin line those with equal costs. We see that with equal costs of market making the optimal prices do not vary with a change in the reservation price, while with different costs substantial changes can be found, although costs have only changed slightly. In the lower panels of reservation price schemes (c) and (d) the negative relationship between the liquidity event and the optimal price nearly vanishes.

The simulation has been conducted as follows: we drew 50,000 random liquidity events, in one case from a uniform distribution between 2.5 and 5, denoted $U(2.5, 5)$ and in the other case from a truncated normal with mean 2.5 and variance .5 with values between 2.5 and 5, denoted $N_{[2.5,5]}(2.5, .5)$. Each of these random liquidity events is assigned to one of the four reservation price schemes with equal probability of .25. To determine the relationship between the liquidity event and the quoted price, we regressed the quoted prices on the liquidity event using the method of ordinary least squares.

In all cases we found a statistically significant negative relationship between the quoted price and the liquidity event, however, the goodness of fit showed very different values. In table 5.2 we report the R^2 for each simulation considered. We see that in all cases with equal costs of market making the R^2 has values above .8, except for a tick size of $d = ^1/_4$. With these results we should expect to find empirical evidence of implicit collusion from observing the behavior of prices.

With different costs of market making, however, things change significantly. In cases

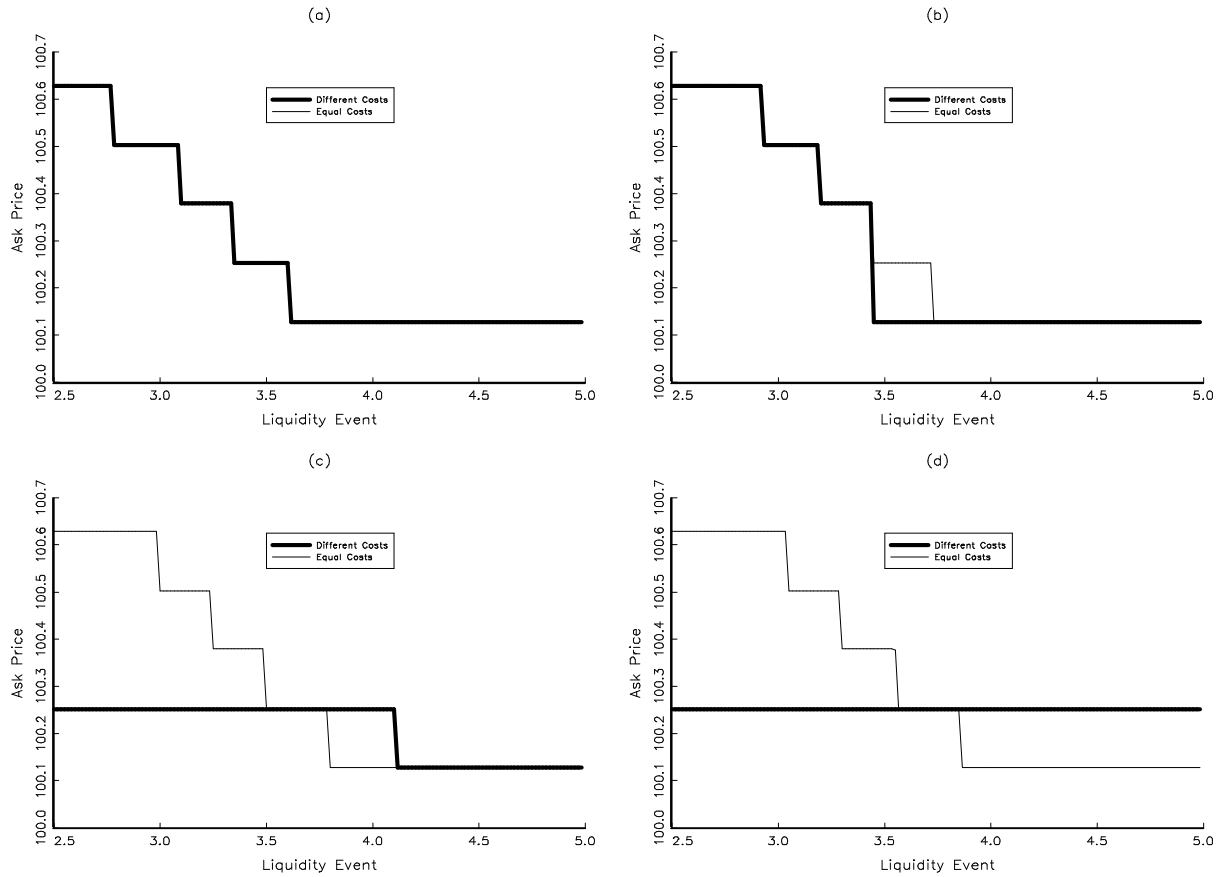


Figure 5.8: Optimal prices with different reservation price schemes and a tick size of $^1/_8$

$L \sim U(2.5, 5)$	equal costs	different costs
d=$^1/_4$	.66	.26
d=$^1/_8$	.83	.46
d=$^1/_{16}$	.91	.39
d=$^1/_{32}$	.92	.37
d=0	.88	.32

$L \sim N_{[2.5,5]}(2.5, .5)$	equal costs	different costs
d=$^1/_4$	.47	.06
d=$^1/_8$	.80	.12
d=$^1/_{16}$	.84	.12
d=$^1/_{32}$	.86	.16
d=0	.86	.18

Table 5.2: R^2 for a regression of the quoted prices on L

where large liquidity events are frequent to occur, which we modeled by assuming the liquidity event to be uniformally distributed, the R^2 reduces to approximately .4. In the more realistic case of only few very large liquidity events, we assumed the liquidity events to be truncated normally distributed, this relationship further weakens to an R^2 only slightly above .1.

Simulations undertaken with other reservation price schemes gave similar results. However, this evidence on a poor negative correlation between the liquidity event and the observed prices is based only on an *ad hoc* analysis. It would be more appropriate to conduct a Monte-Carlo study by simulating the order flow as well as the inventory dynamics of market makers. Our findings derived here, makes it very probable that the results will not give a different evidence.[57]

These results suggest that any empirical investigation is not very likely to give evidence of implicit collusion from observing quoted prices. For the simulation

[57] As we are not interested in time series properties of the quoted prices, we do not have to care about the dynamics in the reservation prices. The important property only is that different prices are quoted for the same liquidity event due to different costs of market making. For this reason no difference can be expected from a simulation of the order flow and inventory dynamics of market makers.

undertaken here, we created optimal conditions that cannot be met empirically. Usually we will not know L_c, hence we have to estimate this parameter. The liquidity event itself cannot directly be observed. As it becomes visible only in the demand of investors, we may use trading volume to approximate the liquidity event. However, trading volume and adverse selection costs are positively correlated, such that we have to eliminate the effects arising from adverse selection costs first. We saw in chapter 4.7.2 that estimating adverse selection costs imposes severe difficulties. Facing all these empirical problems, evidence for or against implicit collusion will be very difficult to find with different and time varying costs of market making.

5.3.3 Imperfect knowledge of the liquidity event

In the previous models it has been assumed that market makers know the realization of the current liquidity event with certainty before quoting their prices. We will now lift this assumption by assuming market makers to receive only a noisy signal of the liquidity event before posting their quotes. We will assume the behavior of market makers still to be perfectly observable, i.e. at the end of the period their signals are revealed to the other market makers.[58] For simplicity, however, we will again assume all market makers to have zero costs of market making. We can later easily combine the results obtained here with those for different costs of market making derived in chapter 5.3.2. To allow a concentration on the effects different signals have on the price setting behavior, let us assume that market makers are alike in all other aspects, i.e. they are equally patient and have the same share of the order flow when quoting the same price.

Each market maker $i \in \mathcal{P}$ receives a private noisy signal $L_i \in [0, \overline{L}]$ of the liquidity event, but not the liquidity event itself. This signal must not be communicated to

[58] KANDORI AND MATSUSHIMA (1998) show that communication is very important to facilitate collusion in situations where players have different information regarding the relevant variables. Experimental evidence from CASON (2000) suggests that without any communication among market makers, spreads are smaller than when they are allowed to communicate. That such an informal communication takes place in markets is reasonable to assume. Market participants frequently express their interpretations of the ongoing activity, which is likely to include information on their inferences of the level of demand, i.e. the liquidity event.

other market makers.[59]

$$L_i = L + \varepsilon_i \tag{5.38}$$

for all $i \in \mathcal{P}$. ε_i denotes a noise term with

$$\begin{aligned} \mu_\varepsilon &= E[\varepsilon_i] = 0, \\ \sigma_\varepsilon^2 &= Var[\varepsilon_i] \geq 0. \end{aligned} \tag{5.39}$$

As L and L_i are bounded, ε_i cannot be independent of L, but it is assumed that the conditional distributions of L and ε_i, for any L_i, are continuous and common knowledge.

The case $\sigma_\varepsilon^2 = 0$ corresponds to a perfect knowledge of the liquidity event as described in chapter 5.3.1 and with $\sigma_\varepsilon^2 = \infty$ the signal is uninformative.[60] The larger σ_ε^2, the less informative the signal is.

Each market maker will determine the expected demand according to the liquidity event he expects given his information, $\mu_i \equiv E[L|L_i]$. We will denote this expected demand by $d(p, \mu_i)$. As the signal and the true liquidity event are in general positively correlated, all properties derived for the demand and expected profits for a known liquidity event can also be applied to the case with only an imperfect knowledge of the liquidity event.[61]

Because μ_i is a conditional mean, although $L_i \in [0, \overline{L}]$ and $L \in [0, \overline{L}]$, it is obvious from distribution theory that μ_i in general cannot take values close to the extremes of 0 and $\overline{L}$,[62] hence we find that with $\underline{\mu} > 0$ and $\overline{\mu} < \overline{L}$ it is $\mu_i \in [\underline{\mu}, \overline{\mu}]$. The

[59] A partial revelation of the signal to other market makers would not change the results. However, with a full revelation of the signals market makers will aggregate their information perfectly and finally agree on the distribution of the liquidity event. They by this mean become homogeneous and the results of chapter 5.3.1 can be applied.

[60] In a mathematically correct sense the variance cannot be infinite as ε_i is bounded. For notational simplicity we here define a complete uninformative signal to have infinite variance.

[61] For this reason we will for simplicity also μ_i denote as the signal received by a market maker. He will base all his decisions on μ_i rather than directly on L_i.

[62] As L and L_i are bounded by assumption, μ_i cannot be close to 0 or $\overline{L}$. Suppose that $\mu_i = \overline{L}$, then the entire probability of the conditional distribution has to be below μ_i. To find that the expected value of this conditional distribution is μ_i then requires the density to

less precise the signal is, the larger (smaller) $\underline{\mu}$ ($\overline{\mu}$) is. If the signal is completely uninformative, we find $\underline{\mu} = \overline{\mu} = E[L]$.

With each market maker receiving a different signal, they will not only differ in their beliefs about the distribution of the liquidity event, but also in their inferences about the signals received by other market makers and their behavior. As the signal cannot be inferred with certainty by other market makers due to the lack of communication at this point of time, each market maker will base his decisions only on his signal and therefore can behave in a different way, i.e. quote different prices.

If all market makers quote different prices, the share of the order flow is not longer fixed but depends on the signal other market makers receive. By making his inferences about the other market makers' behavior, each market maker can calculate the probability λ of receiving the order flow based on his signal:

$$\begin{aligned}(5.40)\quad \lambda(p^i, \mu_i) \;\equiv\;& \text{Prob}\left(p^i < p^j \quad \forall j \in \mathcal{P} | \mu_i\right) \\ & + \phi(\mathcal{Q}) \text{Prob}\left(\forall j \in \mathcal{Q} \subseteq \mathcal{P} : p^i = p^j, \forall j \notin \mathcal{Q} : p^j > p^i | \mu_i\right) \\ =\;& 1 - \text{Prob}\left(\exists j \in \mathcal{P}, i \neq j : p^i > p^j | \mu_i\right),\end{aligned}$$

for all $i \in \mathcal{P}$ and $\sum_{i \in \mathcal{Q}} \phi(\mathcal{Q}) = 1$. Here p^i denotes the price quoted by market maker i. The expected profits per period are then given by

$$(5.41)\qquad \pi\left(p^i, \mu_i\right) = \lambda(p^i, \mu_i)\left(p^i - p^*\right) d(p^i, \mu_i).$$

From definition (5.40) λ obviously is monotonically decreasing in p^i and $1 - \lambda$ can be interpreted as a probability distribution with $1 - \lambda(p^*, \mu_i) = 0$ and $1 - \lambda(\infty, \mu_i) = 1$.

In analogy to the models above, market makers maximize their expected value of future profits:

$$(5.42)\qquad J(\mu_i) = \max_{p^i} \left\{\pi(p^i, \mu_i) + \rho E\left[J(\mu_i)\right]\right\}.$$

degenerate taking positive values only at $\overline{L}$. As the distributions of L and ε_i are continuous such a degeneration can only occur if $\sigma_\varepsilon^2 = 0$. Unless $\sigma_\varepsilon^2 = 0$ we obtain $\mu_i < \overline{L}$. The same argumentation can be used to show that $\mu_i > 0$. The higher the variance,σ_ε^2, the higher the conditional variance of L. Therewith with the distribution not being able take values below (above) 0 $\overline{L}$, μ_i increases (reduces).

To avoid a defection, expected profits from collusion have to exceed those from defection. With p_c^i denoting the collusive price, the expected profits from collusion are given according to (5.42). The expected profits from defection are denoted by $\pi^D(p_D^i, \mu_i)$, where p_D^i denotes the price to be applied under defection. As all market makers face the same costs of market making, the competitive pricing applied afterwards gives expected profits of zero. Hence the constraint under which (5.42) has to be maximized is

$$\pi^C(p_c^i, \mu_i) + \rho E\left[J(\mu_i)\right] \geq \pi^D(p_D^i, \mu_i). \tag{5.43}$$

In the following we will give the intuition for the properties of the collusive equilibrium. A formal proof of these claims is provided in appendix E.

Unlike in the previous models, it is no longer optimal for defecting market makers to quote a price only a fraction below the collusive price. As all market makers receive different signals on the size of the liquidity event, they in general will apply different prices. Hence, by reducing the price only marginally, a market maker cannot expect to receive the entire order flow, instead he will only receive a minor increase in the probability of receiving the order flow. The more he reduces his price, the higher this probability will become. It is therefore in general optimal to quote a price strictly below the collusive price, what reduces the profits from defection and therewith facilitates collusion.

We can rewrite (5.43) as

$$\pi^D(p_D^i, \mu_i) - \pi^C(p_c^i, \mu_i) \leq \rho E\left[J(\mu_i)\right]. \tag{5.44}$$

As before, the profits from defection will increase faster in μ_i than from cooperation, causing the left hand side to increase in μ_i, while the right hand side is a constant. Hence, like in the Dutta-Madhavan model, we find that for small signals this constraint is not binding and prices are determined according to (5.42), what we call the cooperative prices. With the same argumentation as in the Dutta-Madhavan model we find that the cooperative price is increasing in the signal. Hence there

will exist a signal μ_c such that the constraint is fulfilled with equality by quoting the cooperative price.[63] For larger signals the constraint becomes binding and the prices decrease in μ_i to fulfill this constraint.

Unlike in the case of a perfect observable liquidity event, there will be no single signal μ_c at which constraint (5.44) becomes binding and at which the quoted price is highest. If there would exist such a signal, the market maker receiving this signal knows that with probability 1 all other market makers will receive signals implying them to quote a lower price, hence he will make zero expected profits.[64] By lowering the price he increases his probability of quoting the lowest price, hence his expected profits become positive. But now for the liquidity events closest to μ_c the highest price is quoted, for which reason they will also be lowered. This process continuous until in an entire region around μ_c the same price is quoted.

The collusive equilibrium can therewith be characterized as follows:

Theorem 5. *If* $\forall i \in \mathcal{P} : \rho > \rho_0^i = 1 - \frac{\pi(p_c^i, \mu_i)}{E[J(\mu_i)]}$ *a collusive equilibrium exists with the following properties:*

There exist $\underline{\mu} \leq \mu^1 \leq \mu_c \leq \mu^2 \leq \overline{\mu}$ *such that for*

- $\mu_i < \mu^1$ *the collusive price equals the cooperative price and is increasing in* μ_i,
- $\mu^1 \leq \mu_i \leq \mu^2$ *the collusive price remains constant and for*
- $\mu_i > \mu^2$ *the collusive price is decreasing in* μ_i.

If $\exists i \in \mathcal{P} : \rho < \rho_0^i$ *the only equilibrium is to quote competitive prices.*

The condition on ρ in this theorem implies that for small signals the expected profits may not be sufficiently large to ensure a collusive equilibrium.

[63] As all market makers are alike, except for the signal they receive, the constraint becomes binding for all market makers at the same signal. Furthermore, it is obvious that for any given signal the market makers apply the same price.

[64] As all random variables have been assumed to be continuous, the probability of another market maker receiving the same signal is zero.

With the result that the expected profits are increasing in μ_i and noting that $E[J(\mu_i)]$ is a constant, ρ_0^i is decreasing in μ_i. Hence, if $\pi(p_c^i, \underline{\mu})$ is sufficiently small, there exists a μ_0 such that for all $\mu_i < \mu_0$ we find $\rho < \rho_0^i$ and according to theorem 5 competitive pricing will be applied.[65] However, for larger signals the collusive prices will be applied and the larger signal, the less patient market makers have to be.

We find further that compared to a perfect knowledge of the liquidity event, the co-operative price will always be lower. To see this, investigate the first order condition of a profit maximum as easily can be derived from (5.41):[66]

$$\begin{aligned} \frac{\partial \pi(p^i, \mu_i)}{\partial p^i} &= \frac{\partial \lambda\left(p^i, \mu_i\right)}{\partial p^i}(p^i - p^*) d(p^i, \mu_i) \\ &\quad + \lambda\left(p^i, \mu_i\right)\left[d(p^i, \mu_i) + (p^i - p^*)\frac{\partial d(p^i, \mu_i)}{\partial p^i}\right] = 0. \end{aligned} \tag{5.45}$$

In the case of perfect observable liquidity events, the share of the order flow, λ, is fixed to ϕ. Therewith we have $\frac{\partial \lambda\left(p^i, \mu_i\right)}{\partial p^i} = 0$ and the first order condition in (5.37) reduces to

$$\frac{\partial \pi(p^i, \mu_i)}{\partial p^i} = \phi\left[d(p^i, \mu_i) + (p^i - p^*)\frac{\partial d(p^i, \mu_i)}{\partial p^i}\right] = 0. \tag{5.46}$$

Suppose now that the optimal price is the same in both cases. Inserting the term in brackets from (5.46), which has to be equal zero, into (5.45) implies that with the result of $\frac{\partial \lambda\left(p^i, \mu_i\right)}{\partial p^i} < 0$ in the case of imperfect knowledge of the liquidity event, $\frac{\partial \pi(p^i, \mu_i)}{\partial p^i} < 0$. The concavity of the expected profits then suggests that the optimal price has to be strictly below the optimal price in the Dutta-Madhavan model.

The more informative the received signal on the liquidity event is, the more alike will be the inferences of the market makers, hence the effects on λ of applying another price will be smaller as all other market makers follow a similar strategy, i.e. $\frac{\partial \lambda\left(p^i, \mu_i\right)}{\partial p^i} < 0$ increases. Using the same argumentation as above, we see that the first term in (5.45) becomes larger and hence the optimal price increases in the

[65] As for any given signal the same price is quoted, the expected profits will also be equal. Therewith ρ_0^i is the same for all market makers, hence μ_0 is also identical.

[66] As before the present value of future profits is maximized by maximizing the expected profits in each period.

precision of the signal. Therewith we find that the less informative the signal is, the smaller the cooperative price will be for the same signal μ_i.

We know from equation (E.25) of appendix E that[67]

$$\frac{\partial p_c^i(\mu)}{\partial \mu}\bigg|_{\mu=\mu_i} = 2\frac{\frac{\partial \varphi^C(p^i,\mu)}{\partial \mu}\Big|_{p^i=p_c^i}^{\mu=\mu_i}}{-3\,\frac{\partial \varphi^C(p^i,\mu)}{\partial p^i}\Big|_{p^i=p_c^i}^{\mu=\mu_i}} + \frac{1}{3}. \tag{5.47}$$

Denoting σ a measure for the precision of the signal we get

$$\frac{\partial^2 p_c^i(\mu)}{\partial \mu \partial \sigma}\bigg|_{\mu=\mu_i} = 2\frac{-3\,\frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \mu \partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i}\,\frac{\partial \varphi^C(p^i,\mu)}{\partial p^i}\Big|_{p^i=p^*}^{\mu=\mu_i} + 3\,\frac{\partial \varphi^C(p^i,\mu)}{\partial \mu}\Big|_{p^i=p_c^i}^{\mu=\mu_i}\,\frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma}\Big|_{p^i=p^*}^{\mu=\mu_i}}{9\left(\frac{\partial \varphi^C(p^i,\mu)}{\partial p^i}\Big|_{p^i=p_c^i}^{\mu=\mu_i}\right)^2} \tag{5.48}$$

Using first order Taylor series approximations we receive

$$\frac{\partial \varphi^C(p,\mu_i)}{\partial \sigma}\Big|_{p^i=p^*}^{\mu=\mu_i} = \frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i} + \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i}(p^* - p_c^i), \tag{5.49}$$

$$\frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i} = \frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p^*}^{\mu=\mu_i} + \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma}\Big|_{p^i=p^*}^{\mu=\mu_i}(p_c^i - p^*), \tag{5.50}$$

which solve for

$$\begin{aligned}\frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma}\Big|_{p^i=p^*}^{\mu=\mu_i} &= \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma}\Big|_{p=p_c^i}^{\mu=\mu_i} \\ &= \frac{\frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i} - \frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p^*}^{\mu=\mu_i}}{p_c^i - p^*}.\end{aligned} \tag{5.51}$$

We further get from

$$\frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\underline{\mu}} = \frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i} + \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \sigma \partial \mu}\Big|_{p^i=p_c^i}^{\mu=\mu_i}(\underline{\mu} - \mu_i) \tag{5.52}$$

$$\frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i} = \frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\underline{\mu}} + \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \sigma \partial \mu}\Big|_{p^i=p_c^i}^{\mu=\mu_i}(\mu_i - \underline{\mu}). \tag{5.53}$$

[67] We here also use the convention from appendix E that $\varphi^C(p,\mu) = \lambda^C(p,\mu)d(p,\mu)$ and $\varphi^D(p,\mu) = \lambda^D(p,\mu)d(p,\mu)$ for notational simplicity. To distinguish between a variable and its value in these mathematically more sophisticated paragraphs, we denote a derivative of a function $f(x,y)$ with respect to x evaluated at (x_0,y_0) by $\frac{\partial f(x,y)}{\partial x}\Big|_{x=x_0}^{y=y_0}$.

From theorem 5 we know that $p_c^i(\underline{\mu}) = p^*$ as the collusive equilibrium is not sustainable for $\mu_i = \underline{\mu}$. Hence we can substitute p_c^i by p^* in the first term of (5.53).

We can assume that the probability to quote the best price reacts equal sensitive to changes in the precision of the signal, regardless of the signal. As the demand of investors does not depend on the precision of the signals market makers receive, we obtain

$$\frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\bigg|_{p^i=p^*}^{\mu=\mu_i} = \frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\bigg|_{p^i=p^*}^{\mu=\underline{\mu}}. \tag{5.54}$$

Inserting this into (5.53) this transforms to

$$\frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\bigg|_{p^i=p_c^i}^{\mu=\mu_i} = \frac{\partial \varphi^C(p^i,\mu)}{\partial \sigma}\bigg|_{p^i=p_c^i}^{\mu=\mu_i} + \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \sigma \partial \mu}\bigg|_{p^i=p_c^i}^{\mu=\mu_i} (\mu_i - \underline{\mu}). \tag{5.55}$$

Comparing with (5.50) we can derive that

$$\frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \sigma \partial \mu}\bigg|_{p^i=p_c^i}^{\mu=\mu_i} (\mu_i - \underline{\mu}) = \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \sigma \partial p^i}\bigg|_{p^i=p^*}^{\mu=\mu_i} (p_c^i - p^*). \tag{5.56}$$

For $\mu_i > \underline{\mu}$ this becomes

$$\frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \sigma \partial \mu}\bigg|_{p^i=p_c^i}^{\mu=\mu_i} = \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial \sigma \partial p^i}\bigg|_{p^i=p^*}^{\mu=\mu_i} \frac{p_c^i - p^*}{\mu_i - \underline{\mu}}. \tag{5.57}$$

Inserting these relations into (5.48) we receive

$$\frac{\partial^2 p_c^i(\mu)}{\partial \mu \partial \sigma}\bigg|_{\mu=\mu_i} = 2 \frac{\frac{\partial \varphi^C(p^i,\mu)}{\partial \mu}\Big|_{p^i=p_c^i}^{\mu=\mu_i} - \frac{\partial \varphi^C(p^i,\mu)}{\partial p^i}\Big|_{p^i=p^*}^{\mu=\mu_i} \frac{p_c^i - p^*}{\mu_i - \underline{\mu}}}{3\left(\frac{\partial \varphi^C(p^i,\mu)}{\partial p^i}\Big|_{p^i=p_c^i}^{\mu=\mu_i}\right)^2} \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma}\bigg|_{p^i=p^*}^{\mu=\mu_i}. \tag{5.58}$$

As $\frac{\partial \varphi^C(p^i,\mu)}{\partial p^i}\Big|_{p^i=p^*}^{\mu=\mu_i} < 0$, $\frac{\partial \varphi^C(p^i,\mu)}{\partial \mu}\Big|_{p^i=p_c^i}^{\mu=\mu_i} > 0$ and obviously $\frac{p_c^i - p^*}{\mu_i - \underline{\mu}} > 0$ the first term is positive. As has already been stated above, we find that the probability of receiving the order flow is more affected by changing the price the more precise the signal is. As the demand of investors obviously is unaffected by the precision of the signal the market makers receive, we find $\frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma}\Big|_{p^i=p^*}^{\mu=\mu_i} > 0$. This leads to $\frac{\partial^2 p^i}{\partial \mu \partial \sigma}\Big|_{p^i=p_c^i}^{\mu=\mu_i} > 0$ and the cooperative price increases more in the signal the more precise it is.

To explore the case of $\mu_i = \underline{\mu}$ we can apply the rule of de l'Hôpital to the term $\frac{p_c^i - p^*}{\mu_i - \underline{\mu}}$ as the price depends on the signal received and $p^* = p_c^i(\underline{\mu})$:

$$\lim_{\mu_i \to \underline{\mu}} \frac{p_c^i(\mu_i) - p_c^i(\underline{\mu})}{\mu_i - \underline{\mu}} = \lim_{\mu_i \to \underline{\mu}} \left. \frac{\partial p_c^i(\mu)}{\partial \mu} \right|_{\mu=\mu_i} = \left. \frac{\partial p_c^i(\mu)}{\partial \mu} \right|_{\mu=\underline{\mu}}. \tag{5.59}$$

From proposition 3 in appendix E this term is non-negative. As a result (5.58) becomes for $\mu_i = \underline{\mu}$

$$\frac{\partial^2 p_c^i}{\partial \mu_i \partial \sigma} = 2 \frac{\left. \frac{\partial \varphi^C(p^i,\mu)}{\partial \mu} \right|_{p^i=p_c^i}^{\mu=\underline{\mu}} - \left. \frac{\partial \varphi^C(p^i,\mu)}{\partial p^i} \right|_{p^i=p^*}^{\mu=\mu_i} \left. \frac{\partial p_c^i(\mu)}{\partial \mu} \right|_{\mu=\underline{\mu}}}{3 \left(\left. \frac{\partial \varphi^C(p^i,\mu)}{\partial p^i} \right|_{p^i=p^*}^{\mu=\mu_i} \right)^2} \left. \frac{\partial^2 \varphi^C(p^i,\mu)}{\partial p^i \partial \sigma} \right|_{p^i=p^*}^{\mu=\mu_i}. \tag{5.60}$$

With the same arguments as for $\mu_i > \underline{\mu}$ we see that also for $\mu_i = \underline{\mu}$ it is $\frac{\partial^2 p_c^i}{\partial \mu_i \partial \sigma} > 0$.

Because the cooperative price increases less in less precise signals and therewith also cooperative profits increase less, constraint (5.44) will become binding only for larger signals, i.e. μ_c increases with less precise signals. Furthermore, for larger signals the price has to be decreased less in order to fulfill this constraint, hence for $\mu_i > \mu_c$ prices decrease less in μ_i the less precise the signal is. Depending on constraint (5.44), for $\mu_i > \mu_c$ the quoted price will be below or above the price with a more precise signal. Due to the reduced slope for less precise signals, we can expect that prices are higher for large signals.

As with an uninformative signal, i.e. $\sigma_\varepsilon^2 = \infty$, market makers will not base their decisions on this information, they will quote the same price for any signal they receive. We therefore expect the price not to vary with the signal received.

Figure 5.9 illustrates these findings. We see that, compared to the case of a perfect knowledge of the liquidity event, i.e. $\sigma_\varepsilon^2 = 0$, as in the Dutta-Madhavan model, the slopes are smoothed and for medium signals $\mu_i \in [\mu^1, \mu^2]$ the relationship between the price and the signal is entirely flat. Furthermore, for small signals even a collusive equilibrium may not exist. This smoothing of the slopes and the flat region, where the same properties as with competitive pricing are found, will increase the

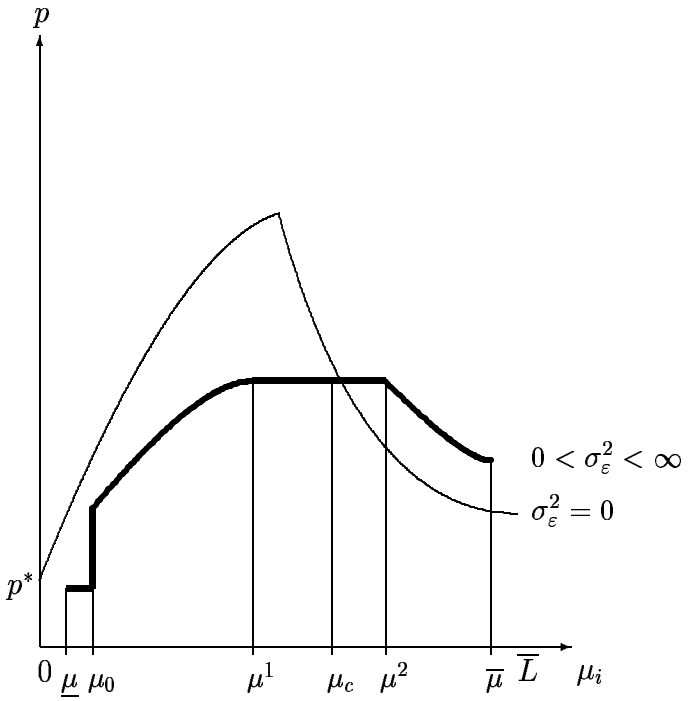


Figure 5.9: Prices quoted by a single market maker depending on his conditional mean

difficulties of proving the existence of implicit collusion empirically by using these characteristics.

When we assume that all distributions are such that a higher signal L_i corresponds to a higher conditional mean μ_i, we can directly transform the above results from a dependence on μ_i to a dependence on L_i. It is obvious that with such a monotone transformation the general properties of the equilibrium do not change. Only the result that the quoted price of the cooperative equilibrium, i.e. the price for $\mu_i < \mu_c$, is always smaller the less precise the signal received, has no longer to be valid. Figure 5.10 shows the effect this transformation from μ_i to L_i has.

It is reasonable to assume that an outside observer does not know the private signal received by each market maker, but in an *ex post* analysis knows the true liquidity event. Using this information, he can determine the conditional distribution of the signals.

Using his conditional distribution of L_i, he can calculate the distribution of quotes

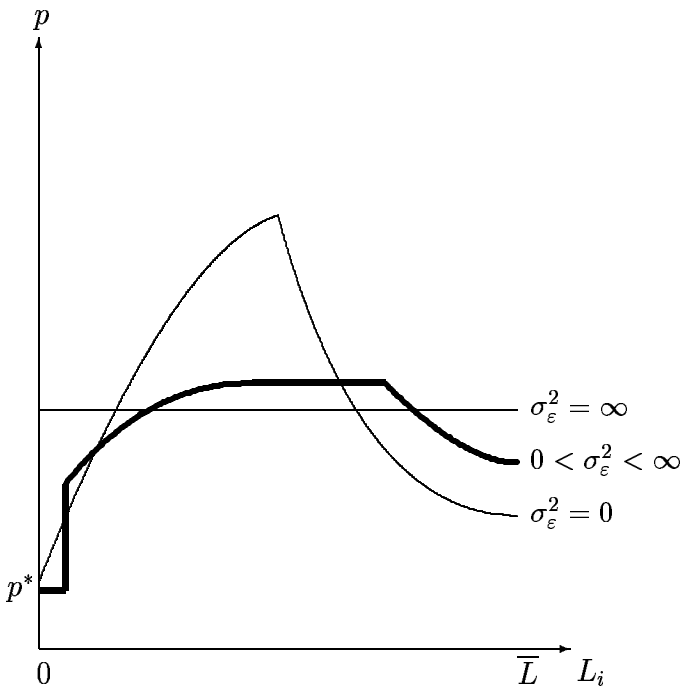


Figure 5.10: Prices quoted by a single market maker depending on his signal

he will observe by weighting the prices determined by the market makers for each possible signal according to the analysis above. With this distribution of quotes he then can determine his expected quotes as well as the expected best available quotes for each state. As we can easily deduct from the properties of the prices in theorem 5 and from figure 5.9, the expected quotes[68] have the property that there exists a liquidity event L^C such that for

- $L < L^C$ expected quotes are increasing in μ_i and for
- $L \geq L^C$ expected quotes are decreasing in μ_i.

For small and large liquidity events expected quotes are above the quotes a market maker will use when receiving a signal of the same magnitude, while for liquidity events closer to L^C expected quotes are lower.

To see this claim, recall that the expected quotes are weighted averages of quotes used by market makers. As it is reasonable to assume that the weight of larger

[68] The expected best available quotes have the same properties as easily can be shown.

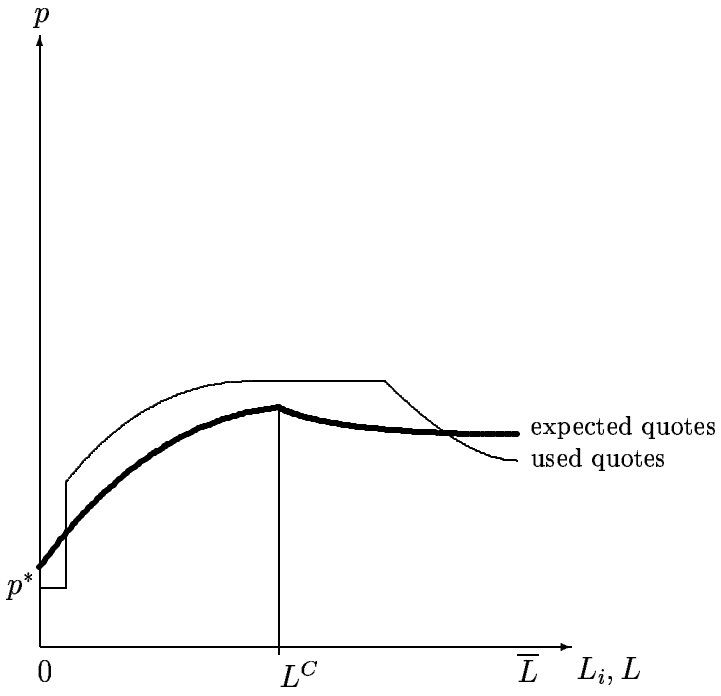


Figure 5.11: Used and expected quotes

signals increases with the true liquidity event, the weights for the associated prices increase.[69] As the used quotes for small signals increase in the signal, the expected quotes will increase, whereas for large signals the used quotes are decreasing and so will the expected quotes. Obviously there will be a liquidity event L^C where this relation changes.

For small and large liquidity events the main weights will be on prices larger than those used by a market maker receiving a signal of the same magnitude, hence the expected quotes will be higher, the opposite is true for medium states. Here the main weight will be on lower prices, hence the expected quote will be lower.

We also see that the more precise the signal received by market makers is, the closer expected and actually applied quotes are, because the conditional distribution of μ_i has a smaller variation and hence prices not close to L only have very small weights.

Figure 5.11 illustrates this finding. The expected quotes are a further smooth of

[69] If the weights of larger signals do not monotonically increase in the true liquidity event, in general nothing can be said about the relation of the true liquidity event and the quoted price.

the actual used quotes. Although the result derived for the expected quotes looks very similar to the result obtained from the Dutta-Madhavan model, it is not only much more smoothed, but also represents only the expected quote rather than the actual used quote. For every liquidity event there exists an entire distribution of quotes, so that even large deviations from the expected quote in small samples are not sufficient to assume a violation of these results.

We may summarize our findings that imperfect knowledge of the liquidity event facilitates implicit collusion between market makers, especially for large signals. On the other hand, except for large and small signals, the quoted price will be lower the less precise the received signal is and therewith trading costs are reduced.

Having only a smoothed distribution of quotes as a result, will also impose additional difficulties in finding empirical evidence for implicit collusion when not knowing the signals received by market makers but only the true liquidity event.

The more precise the signals are, the less smooth the result is and the smaller the variance of the distribution of used quotes for an outside observer.

Therefore, the precision of the signal the market makers receive will have an important role in analyzing and interpreting any empirical result. Nevertheless we can deduct that imperfect knowledge of the liquidity event by market makers, which is the more reasonable case, aggravates the empirical detection of implicit collusion by smoothing the negative relationship between the quoted price and large liquidity events.

5.3.4 Imperfect observable behavior of market makers

We will now use the same model as before, but relax the assumption that the behavior of each market maker can perfectly be observed by the other market makers, i.e. they become to know whether the quote a market maker publishes is motivated by collusion or defection. It is more reasonable to assume that all market makers

can observe quoted prices, the public outcome, but not the motivation for these outcomes, or equivalently the private signal of a market maker. Hence we have to consider equilibria in trigger strategies.

The probability that the other market makers interpret a certain price as defective depends not only on the price quoted, but also on the signals the other market makers receive. When receiving a signal corresponding to a low price, the probability will be smaller than when receiving a signal corresponding to a high price. A market maker has to infer these probabilities from the signal he receives and make conjectures about the distribution of signals the other market makers receive.[70] hence we denote $\gamma(p_c^i, \mu_i)$ as the probability, given a certain signal μ_i, that the trigger price is exceeded, i.e. $p_c^i > \widehat{p}$, and therewith his price is regarded as collusive. The time of punishment the other market makers apply in general will also depend on the signal they receive and hence they will differ between market makers and the defecting market maker has to infer the maximum length from his signal, which we denote $\widehat{T}(\mu_i)$.

Using the results from chapter 5.1.6 we can easily determine the expected payoffs as

$$J^i(\mu_i) = \pi^i(p^i, \mu_i) + \rho\left(\gamma(p^i, \mu_i)E[J^i(\mu_i)] + (1 - \gamma(p^i, \mu_i))\rho^{\widehat{T}(\mu_i)}E[J^i(\mu_i)]\right), \tag{5.61}$$

which is maximized using the optimal price p_c^i. In the same manner each market maker $i \in \mathcal{P}$ chooses his individual trigger price $\widehat{p}_i$ at which he punishes another market maker for $\widehat{T}^i$ trading periods.

To avoid defection, we again have to apply the restriction that the expected profits from collusion exceed those from defection, i.e.

$$\begin{aligned}
&\pi^i(p_c^i, \mu_i) + \rho\left(\gamma(p_c^i, \mu_i)E[J^i(\mu_i)] + (1 - \gamma(p_c^i, \mu_i))\rho^{\widehat{T}(\mu_i)}E[J^i(\mu_i)]\right) \\
&\geq \pi^i(p_D^i, \mu_i) + \rho\left(\gamma(p_D^i, \mu_i)E[J^i(\mu_i)] + (1 - \gamma(p_D^i, \mu_i))\rho^{\widehat{T}(\mu_i)}E[J^i(\mu_i)]\right).
\end{aligned}$$

[70] We can allow for a communication between market makers adding another source of information on the signal received by a market maker. As long as this communication does not allow a perfect revelation of the received signal the results in general will not be affected.

Rearranging this relation gives

$$(5.62) \qquad \pi^i(p_D^i, \mu_i) - \pi^i(p_c^i, \mu_i) \leq \left[\gamma(p_c^i, \mu_i) - \gamma(p_D^i, \mu_i)\right](1 - \rho^{\widehat{T}(\mu_i)})\rho E[J^i(\mu_i)].$$

The in chapter 5.3.3 considered case of perfect observable behavior corresponds to $\gamma(p_c^i, \mu_i) = 1$ and $\gamma(p_D^i, \mu_i) = 0$. With an infinite punishment, i.e. $\widehat{T}(\mu_i) = \infty$, this constraint is reduced to that in chapter 5.3.3.

If at the other extreme, defection cannot be detected, we find $\gamma(p_c^i, \mu_i) = \gamma(p_D^i, \mu_i) = 1$ and therewith $\pi^i(p_D^i, \mu_i) \leq \pi^i(p_c^i, \mu_i)$. As obviously $\pi^i(p_D^i, \mu_i) \geq \pi^i(p_c^i, \mu_i)$, the only solution is $\pi^i(p_D^i, \mu_i) = \pi^i(p_c^i, \mu_i)$, implying $p_c^i = p_D^i$. From proposition 6 in appendix E we know that the optimal defective price is strictly below the cooperative price as long as $\frac{\partial \lambda^D(p_D^i, \mu_i)}{\partial p_D^i} < \frac{\partial \lambda^C(p_c^i, \mu_i)}{\partial p_c^i}$. The requirement for both prices to be equal is $\frac{\partial \lambda^D(p^i, \mu_i)}{\partial p_D^i} = \frac{\partial \lambda^C(p^i, \mu_i)}{\partial p_c^i}$, which is reasonable for $p_c^i = p_D^i = p^*$ as has been pointed out in the proof of proposition 7 in appendix E as equation (E.33), i.e. competitive pricing. If defection cannot be detected, we therewith found that the only equilibrium is to apply competitive prices.

As $\gamma(p, \mu_i) \in [0, 1]$ and $\gamma(p_c^i, \mu_i) \geq \gamma(p_D^i, \mu_i)$, in general we find $0 \leq \gamma(p_c^i, \mu_i) - \gamma(p_D^i, \mu_i) \leq 1$, hence the right side of (5.62) is smaller than with perfect observable behavior and the constraint becomes more binding the less well defective behavior can be detected. A first consequence of imperfect observable behavior therewith is that μ_c reduces, hence prices for signals close to μ_c will decrease. For small and large signals, when relatively small prices are chosen in the above model, we see that $\gamma(p_c^i, \mu_i)$ approaches zero, hence it is optimal to quote a higher price to avoid a punishment of the other market makers receiving signals implying them to belief this market maker to defect.

We therewith find prices are higher for small and large signals and smaller for medium signals. The quoted prices of an individual market maker are further smoothed and the negative relationship between the signal, and hence also the true liquidity event, and the price will become even less visible.

5.3.5 Critical assessment of the results

In this section we showed that implicit collusion between market makers to quote higher ask prices are sustainable under very general conditions, even for frequently traded assets with a large number of market makers. We found a positive relation between the size of the liquidity event and the ask price for small and a negative relationship for large liquidity events. As we assumed the bid prices to behave symmetrically, we can easily see that the same results hold for the quoted spread.

We also showed that under very restrictive assumptions - equal costs of market making, knowledge of the size of the liquidity event and perfect observable behavior of market makers - these properties of the spread may be used to find empirical evidence for or against implicit collusion from market data. Whereas upon relaxing these assumptions, especially upon introducing different and time varying costs of market making, these properties were shown to become much less significant. Introducing noisy observations of the liquidity event and imperfect observable behavior of market makers smoothed the properties, aggravating empirical conclusions. It is obvious that combining all these generalizations into a realistic framework, will make it virtually impossible find any direct empirical evidence for or against the presence of implicit collusion in the price setting behavior.

The liquidity event cannot directly be observed empirically, that is why we have to use an estimate of this variable. As the liquidity event determines the level of demand, which we observe as trading volume, the appropriate estimate should be trading volume. An additional problem arises in empirical investigations. If market makers face adverse selection costs, which are positively related to trading volume as pointed out in chapter 4.4.1, we should expect to find the spread to be increasing in trading volume. In an empirical investigation we therefore have first to estimate the adverse selection component of the spread and analyze the remaining part for evidence of implicit collusion. As has been pointed out in chapter 4.7, all known estimates of spread components are very sensitive to the underlying model and

therefore this estimation will be very difficult to conduct such that the results can be relied on. With these severe empirical and theoretical problems we cannot expect to find any significant and convincing direct evidence of implicit collusion.

For this reason, the above derived results are a purely theoretical analysis which shows that noncompetitive spreads may be the outcome of noncooperative behavior even under very general conditions. In chapter 5.6 we will address some possibilities to give evidence of implicit collusion indirectly through the behavior of market makers. In chapter 6 we consider incentives and measures to reduce the effects arising from implicit collusion.

5.4 Price matching arrangements

Market makers not only compete for order flows through quoting prices, they also use *order preferencing arrangements* and *price matching arrangements*. These arrangements are frequently used on the NASDAQ and other dealer markets as has been pointed out in chapter 3. Order preferencing arrangements are agreements between a market maker and a broker to route the entire order flow to him, provided he quotes the best available price. Price matching arrangements, which we consider here, go a step further by requiring the broker to route his entire order flow to this market maker, regardless of his current quote. In turn the market maker guarantees to execute these orders at the best available price, even if he currently quotes a less favorable price.[71]

Suppose that a fraction $0 \leq \theta \leq 1$ of the entire order flow is preferenced, the remaining order flow is divided as before, where the share of the order flow when quoting the best available price is ϕ^i for market maker $i \in \mathcal{P}$. Each market maker receives a share of θ^i of the preferenced order flow, such that $\sum_{i \in \mathcal{P}} \theta^i = \theta$. The

[71] A more complete modeling of price matching arrangements and its effects on the price setting behavior, although not addressing implicit collusion, is given in KANDEL AND MARX (1999a).

share of the order flow any market maker receives is therefore given by

$$\phi^{i,*} = \theta^i + (1-\theta)\phi^i. \tag{5.63}$$

We will now investigate the collusive behavior following DUTTA AND MADHAVAN (1997) assuming all market makers to have zero costs of market making and a perfect knowledge of the current liquidity event. Replacing ϕ^i by $\phi^{i,*}$ in the more general models in chapters 5.3.2 - 5.3.4 will give similar results as easily can be seen.

We can rewrite the performance function as

$$J^i(L) = \max_{p^i}\{\phi^{i,*}(p^i - p^*)d^i(p^i, L_t) + \rho^i E[J(L)]\}. \tag{5.64}$$

The incentive constraint becomes

$$\phi^{i,*}(p_c^i - p^*)d^i(p_c^i, L_t) + \rho^i E[J(L)] \geq [\theta^i + (1-\theta)](p_c^i - p^*)d^i(p_c^i, L_t), \tag{5.65}$$

which solves for

$$(p_c^i - p^*)d^i(p_c^i, L_t) \leq \frac{\rho^i}{(1-\theta)(1-\phi^i)}E[J^i(L)]. \tag{5.66}$$

The right hand side of (5.65) represents the expected profits from a defection. The market maker receives his share of the preferenced order flow and the entire unpreferenced order flow. Regardless of his quotes he will not receive a share of the order flow preferenced to other market makers, as they will always apply the same price as he does. Therewith preferencing reduces the profits from defection and the constraint becomes less binding. As can be seen from (5.66) the right hand side is increasing in the share of preferenced order flow, such that L_c increases and the collusive price is higher for $L > L_c$. As the cooperative price is not affected by the share of the order flow, for $L < L_c$ the optimal prices are the same with and without price matching arrangements.

We can further show by inserting (5.66) into (5.64) that

$$\begin{aligned} J^i(L) &= \phi^{i,*}(p - p^*)d^i(p, L_t) + \rho^i E[J(L)] \\ &\leq \rho^i E[J^i(L)]\left[\frac{\theta^i + (1-\theta)\phi^i}{(1-\theta)(1-\phi^i)} + 1\right] \\ &\leq \frac{1}{(1-\theta)(1-\phi^i)}\rho^i E[J^i(L)]. \end{aligned} \tag{5.67}$$

Taking expectations on both sides and solving for ρ^i we get[72]

$$\rho^i \geq \rho_0^i = (1-\theta)(1-\phi^i), \tag{5.68}$$

which is decreasing in the share of preferenced order flow, θ. We therewith have shown that price matching arrangements facilitate collusion and increase the collusive price as competition is reduced to the unpreferenced order flow.

Typically market makers pay for the order flow received by a broker, payments between USD .01 and USD .02 per traded share have been reported to be common for NASDAQ stock according to NASD, Inc. (1991). But not only cash payments, also other forms of compensation are frequently found, e.g. reciprocal arrangements between broker-dealers, research reports delivered by market makers or the clearing of trades. All these activities impose costs on the market maker the broker saves, which are outweighed by the benefits from receiving a larger share of the order flow and sustaining the collusive equilibrium more easily. As these costs are reasonably equal for all market makers, an inclusion of such costs in our model would not change the results as easily can be demonstrated.

5.5 Market making with multiple assets

The previous sections investigated the existence and properties of a collusive equilibrium in the price setting behavior of market makers. We therefore assumed that only a single asset is considered. However, in reality a very large number of assets is traded on a market. Furthermore, on markets like the NASDAQ a large number of market makers is admitted to the market and they are all free to register for any asset.

The absence of any entry barriers suggests that in cases where implicit collusion gives rise to excess profits for market makers, we should observe market entries, i.e. registrations for this asset, by some market makers, deteriorating the profits made

[72] As in chapter 5.3.1 it is easy to demonstrate that $E[J^i(L)] > 0$.

by incumbents. If the registration can be withdrawn within a very short period of time at nearly no costs, like on the NASDAQ, hit-and-run competition should strengthen the threat. This threat of new market entries should prevent market makers from executing their market power through implicit collusion. We therefore should observe the market to be contestable, i.e. prices should be set competitively.[73] In this section we will investigate, whether the threat of market entry is credible and under which conditions market makers can execute their market power without fearing market entries.

Market makers usually are registered for more than a single asset and face the same competitors in several markets. On a specialist market like the NYSE, market makers do not directly compete with each other, but we saw in chapter 4.6.2 that assets trading as substitutes may enhance indirect competition and reduce market power. Therefore we can reasonably argue that market makers compete in several markets, i.e. they have *multimarket contacts*. We will consider the implications of multimarket contact to sustain implicit collusion in chapter 5.5.1.

Another question arising in multi-asset markets already pointed out above is, why market makers do not enter all markets if implicit collusion enables them to earn excess profits? If they are free to enter any market, we find that most market makers are only registered for a small fraction of assets. For example, on the NASDAQ market makers are on average only registered for about 100 of the more than 5000 assets trading. In the second section we will therefore consider the issue of market entry that has not previously been considered in the literature.

5.5.1 Implicit collusion with multimarket contact

We assume an economy with $K > 1$ assets. Each market maker $i \in \mathcal{P}$ is registered for each of these K assets.[74] The relevant payoff function J^i for market maker i is

[73] See BAUMOL ET AL. (1988) or TIROLE (1988) for an overview of the theory of contestable markets.

[74] We could without changing the arguments also assume that market makers only meet in $K' < K$ markets and only a subset of $\mathcal{Q} \subset \mathcal{P}$ of market makers meet at all. Adapting notation

supposed to be the sum of the payoff functions in the respective markets, J_k^i:

$$J^i(L) = \sum_{k=1}^{K} J_k^i(L_k), \tag{5.69}$$

where $L = (L_1, L_2, \ldots, L_K)$ denotes the vector of liquidity events for each single asset k and $J_k^i(L_k)$ is determined according to (5.23). We assume for simplicity that the market share of each market maker is the same in each market, i.e. ϕ^i, and the market makers face no costs of market making. Using the same argumentation as in chapter 5.3.1, we get the equivalent incentive constraint of (5.27) for the sustainability of the collusive equilibrium by inserting from (5.69):

$$\sum_{k=1}^{K} J_k^i(L_k) \leq \frac{\rho^i}{1-\phi^i} \sum_{k=1}^{K} E[J_k^i(L_k)], \tag{5.70}$$

which solves for each $i \in \mathcal{P}$ and each K markets to be

$$\rho^i \geq \rho_0^i = 1 - \phi^i, \tag{5.71}$$

by taking expectations on both sides. This result is identical to that derived in the Dutta-Madhavan model with only a single asset, hence multimarket contact does not facilitate collusion. Bernheim and Whinston (1990), who first derived this result, called it an *irrelevance result*. As the payoffs from each asset are included linearly into the total payoff, the optimal prices also do not change upon multimarket contact.

The intuition for this result is straightforward: condition (5.71) requires a market maker either to collude in all or in no market. Therewith multimarket contact only changes the scale of payoffs in (5.69), but not the considerations to be made. It can be noticed that the linearity of payoffs and the assumption of equal market shares makes the consideration in the respective markets independent.

By relaxing these assumptions, Bernheim and Whinston (1990) show that e.g. with different market shares or different costs of market making, collusion is facilitated. There exists a $\rho_0^{i,*} < \rho_0^i$ such that for all $\rho^i \geq \rho_0^{i,*}$ a collusive equilibrium

will give the same results.

exists. More formally, SPAGNOLO (1999) shows that collusion is facilitated if the payoff function (5.69) is concave in the payoffs of each market. This concavity makes the decisions in the single markets interdependent. A defection in one market can give rise to a punishment in several other markets and therewith reduce the payoffs from defection more, which facilitates to sustain collusion.

The concavity of the payoff function is reasonable not only if we assume different shares of the order flows in the market, but also through the saturation effect and risk aversion. Empirical evidence supports the assumption of a concave payoff function.

If we assume that market makers cannot observe their competitors' behavior perfectly as in chapter 5.3.4, we had derived that prices, except for very small and very large signals, are lower than with perfect monitoring. However, if market makers meet in a sufficiently large number of markets, MATSUSHIMA (1998) shows that the same outcomes as with perfect observable behavior can be sustained. Hence, multimarket contact facilitates implicit collusion also in this case.

We find that multimarket contacts facilitate implicit collusion without changing the properties of the collusive equilibrium as derived in chapter 5.3.

It may be worth noting that market makers in most cases form part of larger securities companies. These firms do not only meet in market making, but typically also in brokerage, investment banking, and portfolio management. These additional contacts outside market making may further facilitate collusion.

Thus far we assumed the contacts to be exogenously given. In the following section we will now investigate which assets a market maker should choose to establish multimarket contacts.

5.5.2 The equilibrium control structure with two assets

On multiasset markets like the NASDAQ, it is reasonable to assume the entity of market makers admitted to the exchange to be fixed and given. The entry barrier for

new admittance to the market is very high, while once admitted to the market they are free to register for any asset without meaningful entry barriers. As we find that market makers do not register for all assets, there is the possibility of a market maker entering the market for an asset if he finds this profitable. In the absence of any meaningful restrictions on the withdrawal of registrations, like on the NASDAQ, the theory of contestable markets suggests that hit-and-run competition should prevent market makers from executing their market power, hence implicit collusion should not be sustainable in this framework with potential entry due to the presence of market makers with a time horizon of only a single period, i.e. $\rho^i = 0$.

Only recently VAN WEGBERG AND VAN WITTELOOSTUIJN (1991), VAN WEGBERG AND VAN WITTELOOSTUIJN (1992a) and VAN WEGBERG AND VAN WITTELOOSTUIJN (1992b) developed models addressing the framework mentioned above. They concentrate on the decision of a company to enter a new market. Whereas VAN WEGBERG AND VAN WITTELOOSTUIJN (1992a) concentrate on the hit-and-run competition in such a framework, we will develop a different model here that allows for strategic behavior of market makers.

As we assume that no resources are needed to act as market maker, like capacities in industrial production, we can neglect any aspects arising from the transfer of resources from one market to another.[75]

We will now derive a model to determine under which conditions a market maker enters a market and for which assets he will be registered in equilibrium.

The basic assumptions underlying the model are those also widely applied in the literature:

Assumption 1. *We consider an economy where $N > 1$ assets are traded in an unlimited number of trading rounds at discrete, predetermined, and equally spaced points of time solely by the use of registered market makers.*

[75] See PHLIPS (1995) for models taking into account effects arising from the existence of capacities for implicit collusion.

Assumption 2. *Market makers face no costs of market making and are equally patient.*

Assumption 3. *The number of market makers admitted to the entire market is finite and fixed. Market makers are free to register as market maker for any asset without restrictions.*[76] *Registrations can be made at the beginning of every trading round in a simultaneous choice by all market makers. Conditioning the own decision on the behavior of other market makers is not allowed and the withdrawal of registrations is prohibited.*

Assumption 4. *Market makers have market power arising from implicit collusion. Implicit collusion is only allowed between market makers controlling the same asset,*[77] *implicit collusion across assets is only allowed if all colluding market makers have the same control structure. Market power reduces with the number of registered market makers. This reduction is larger the more market makers are registered for the asset.*

A few remarks concerning these assumptions seem to be useful. Assumption 1 is a common approach in most models of market microstructure. By requiring all trades to be conducted through market makers, we concentrate our analysis on the execution of market power by market makers, rather than effects arising from outside competition.

Although assumption 2 seems to be very restrictive, the absence of costs of market making allows us to concentrate on the effect registrations have on profits. Including costs of market making would give no further substantial insights into our model. In chapter 5.5.3 we waive the assumption of equally patient market makers.

The total number of market makers admitted to the entire market is found to be relatively stable on virtually all markets, including the NASDAQ, so that at least

[76] A market maker is said to be registered as market maker for an asset if he is allowed to act as market maker for this asset.

[77] An asset is said to be *controlled* by a market maker if he is registered as market maker for this asset. The entity of all assets controlled by a market maker is called his *control structure*.

in the short run assumption 3 is realistic. On the NASDAQ market makers can withdraw their registration every trading day, in the generalizations we will also consider the possibility to withdraw the registration after a given period of time.

That market power can arise from implicit collusion, as supposed in assumption 4, has been shown in chapter 5.3. We also saw that with an increasing number of market makers, i.e. decreasing shares of the order flow, implicit collusion is aggravated and prices for larger liquidity events will be lower, i.e. market power decreases. In chapter 5.5.1 we gave evidence that multimarket contacts facilitate implicit collusion further.

With these assumptions it is now possible to construct the basic model. In order to facilitate the analysis, it is necessary to restrict the model more than implied by the basic assumptions stated above. Assume that $M \geq 2$ market makers are initially restricted to choose between one of two exogenously given control structures. Initially the control structures are allocated among market makers such that an equilibrium has been established. Each control structure consists of a single asset, hence we consider an economy with $N = 2$ assets.

Every market maker belongs to one group of market makers, each controlling one of the two assets. As all market makers are alike, their decisions will also be alike and only the behavior of a single market maker out of each group has to be analyzed.

Assuming the initial control structure to be an equilibrium, implies that market makers in both groups receive the same expected profits. To see this assume that market makers in group 1 make larger profits. In this case market makers of group 2 would prefer to withdraw their registration from this asset and register for the other asset, hence it is not an equilibrium. To achieve an equilibrium, the number of market makers registered for the first asset will increase, reducing market power and hence profits for this asset, and the number of market makers registered for the second asset will decrease, increasing profits, as stated in assumption 4. This process will continue until both profits equal.

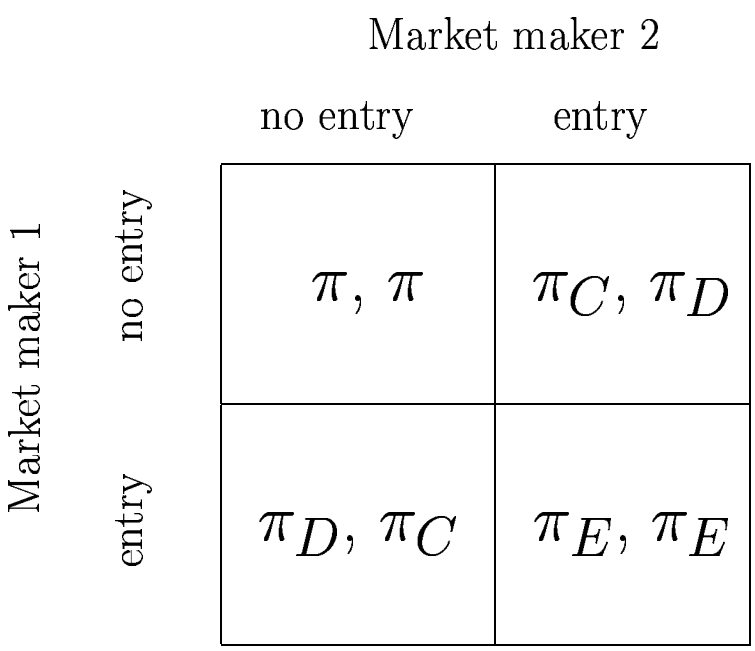


Figure 5.12: Payoff matrix of the entry game

Now let the market makers not only be restricted to one of these two control structures, but allow them also to register for the other asset without enabling to withdraw their registration of their initial asset. If a market maker registers as market maker for the other asset, this is called to "*enter his opponent's market*".

Whether a market maker enters his opponent's market depends on the profits he expects and also on the behavior of the other market maker. Figure 5.12 shows the profits per trading round, depending on the decisions of the market makers.

The symmetry of profits can be explained as follows: For both market makers not entering their opponent's market it is the result of the initial control structure being an equilibrium as pointed out above. For both market makers entering their opponent's market, they have the same control structure and as all market makers are alike, they make the same profits. If one market maker enters his opponent's market while the other does not, market power is reduced by the same degree, regardless which market maker enters and we assume the profits to be equal in both cases. In chapter 5.5.3 we will also consider asymmetric payoffs and show that the

results remain valid in this more general case.

If a market maker enters his opponent's market, assumption 4 states that in this case market power and therewith profits are reduced, hence we find that $\pi > \pi_C$ and $\pi_D > \pi_E$.

This simple model now enables us to analyze the equilibrium outcome of the control structure.

Depending on profits, different sets of Nash equilibria in the above game can occur:

1. $\pi > \pi_D$: The unique Nash equilibrium is that both market makers do not enter their opponent's market.

2. $\pi < \pi_D$ and $\pi_C < \pi_E$: The unique Nash equilibrium is that both market makers enter their opponent's market.

3. $\pi < \pi_D$ and $\pi_C > \pi_E$: Three Nash equilibria exist, where two are characterized by one market maker entering the market, while the other does not, and the third equilibrium is in mixed strategies for both market makers.

If upon entering the opponent's market, market power in the entered market reduces significantly, additional profits made from entering this market are only small. Furthermore, as market makers are not allowed to collude in the price setting behavior across assets if only a single market maker enters the market of his opponent, increased indirect competition from reduced trading costs in the market entered can cause to decrease the market power in the entrant's market substantially. The loss of profits in this market may outweigh additional profits from entering the other market, causing overall profits to fall, even if the other market maker does not enter his opponent's market, i.e. we find $\pi > \pi_D$. Such a situation can be found if the two assets trade as very close substitutes, i.e. indirect competition causes profits to fall, as shown in chapter 4.5.2.

As with $\pi > \pi_D$ we find that the only Nash equilibrium is not to enter their opponent's markets, the assets are controlled by different market makers. For the remainder it is assumed that market power does not decrease so far that $\pi > \pi_D$ and we do not consider this equilibrium further.

If increased indirect competition does not outweigh the benefits from entering an additional market, i.e. $\pi < \pi_D$, market entry of at least one market maker is profitable. The incumbent's profits decrease to π_C if he does not enter his opponent's market as market power decreases. If he also enters his opponent's market, market power in this market is also reduced. Furthermore indirect competition between assets causes market power in the other market to decrease further. But as both market makers have the same control structure, assumption 4 allows them to collude across assets, causing market power to increase. However, the larger number of market makers in the market limits market power. Here we assume that implicit collusion outweighs the effect of indirect competition such that $\pi_C < \pi_E$. Therewith we find that the only equilibrium for both market makers is to enter their opponent's markets.

Both market makers entering their opponent's markets has the largest profits only if $\pi_E > \pi$, otherwise the equilibrium gives both market makers a less preferred outcome. By allowing market makers to collude across assets, market power can increase from joint control of both assets. Using the results from chapter 4.5.2, we can state that if assets trade as substitutes, market power is increased from controlling both assets. We assume these effects to outweigh the reduction in market power due to the larger number of market makers being registered. Hence market makers receive higher profits from controlling both assets, i.e. $\pi_E > \pi$. If they trade as complements, market power decreases when controlling both assets, i.e. $\pi > \pi_E$.

If assets trade as complements, it would be more profitable not to enter the opponent's market, but with $\pi < \pi_D$ this strategy is not a Nash equilibrium.

Thus far only profits from a single trading round have been considered, but as stated

in assumption 1, an infinite number of trading rounds is conducted. Suppose now that a market maker enters his opponent's market in the current period, while the other does not. He receives a profit of $\pi_D > \pi$ and his opponent of $\pi_C < \pi$. As the decision to enter the opponent's market is irrevocable by assumption 3, in the next period the market maker not having entered his opponent's market will also enter because $\pi_E > \pi_C$, and both will receive π_E in each remaining trading round. An entry will only be profitable if the present value of profits in each trading round is increased from entering the market. Hence, to avoid a market maker to enter his opponent's market the following condition has to be fulfilled:

$$\frac{1}{1-\rho}\pi \geq \pi_D + \frac{\rho}{1-\rho}\pi_E, \tag{5.72}$$

where $0 \leq \rho \leq 1$ denotes the discount factor for future profits. This discount factor represents the patience of the market makers and is for simplicity assumed to be equal for all market makers.

It can easily be shown that for $\pi_E > \pi$ this condition is always violated, hence for substitutes both market makers always will enter their opponent's market. For $\pi > \pi_E$, i.e. complements, this in general will not be true. If market makers are sufficiently patient, i.e. if

$$\rho \geq \rho_0 = \frac{\pi_D - \pi}{\pi_D - \pi_E}, \tag{5.73}$$

we see that (5.72) is always fulfilled and both market makers will not enter their opponent's market.

therewith we find that if $\pi < \pi_D$ and market makers are sufficiently patient, they will not enter their opponent's market if assets are complements. If assets are substitutes, market makers always will enter their opponent's market.

given this result, it is obvious which control structure should be expected:

Theorem 6. *Provided market makers are sufficiently patient, assets that trade as substitutes will be controlled by the same market makers, whereas assets that trade as complements will be controlled by different market makers.*

We will now briefly consider strategies of incumbent market makers to avoid market entries of their opponents. Equation (5.72) can be transformed into

$$\pi_D \leq \frac{\pi - \rho\pi_E}{1-\rho}. \tag{5.74}$$

It can be seen that to avoid market entries, market makers have either to decrease π_D or to increase $\pi - \rho\pi_E$.

If incumbent market makers have the ability to coordinate their price setting such that in case of a market maker entering their market, they quote more competitive prices than they could by executing their market power, π_D is reduced. If this behavior is known to possible entrants it may turn out that condition (5.74) is fulfilled and no entries will occur, although with a full execution of market power an entry would have been profitable.

Another option to avoid entries is to increase π. Relaxing the assumption that market makers cannot collude across assets if they do not have the same control structure, would allow market makers to make larger profits in their respective markets. Although π_D in this case also increases, it cannot increase more than π, hence entering the opponent's market may not be profitable.

The final option to reduce π_E, e.g. by refusing to collude with entrants, is not credible as it also reduces the incumbents profits in the remaining trading rounds. It is not a subgame perfect equilibrium.

5.5.3 Generalizations of the model

In this section we will now waive several assumptions made at the beginning and show that theorem 6 remains valid under much more general conditions.

Market makers with different patience. We suppose now that market makers have different discount factors ρ^i, i.e. they have different patience. Regardless of the discount factors, condition (5.72) will always be violated for assets trading as

substitutes and both market makers will enter their opponents' markets. With assets trading as complements condition (5.73) has to be fulfilled for both market makers, i.e.

$$\forall i = 1, 2 : \rho^i \geq \rho_0 = \frac{\pi_D - \pi}{\pi_D - \pi_E}, \tag{5.75}$$

to avoid market entries. Therewith both market makers have to be sufficiently patient, if only one of the market makers is not sufficiently patient, they will both enter their opponents' market.

hence theorem 6 holds with both market makers being sufficiently patient.[78]

Asymmetric payoffs. We made the additional assumption that payoffs in the stage game are symmetric. This assumption will not be realistic because generally markets have different sizes, i.e. trading volume differs. We can furthermore assume that market makers have a different relative positions in their respective initial markets, e.g. as the result of preferencing arrangements. This makes a situation with asymmetric payoffs more realistic. The multimarket contact of market makers facilitates implicit collusion among them despite their different sizes, as we know from BERNHEIM AND WHINSTON (1990) and SPAGNOLO (1999). We therefore can reasonably assume implicit collusion to have similar effects on the profits as described above.

Let the profits of the first market maker be denoted as before and for the second market maker a prime is added to his respective profits. We further assume that the ordering of payoffs is unchanged. Therewith condition (5.72) is violated for assets trading as substitutes and both market makers enter their opponents' markets.

[78] The possibility to withdraw the registration as market maker for an asset to be considered below, would enlarge the strategy space by allowing to trade profits across time. The impatient player would enter his opponent's market immediately, but the patient player will not react by also entering. After a given time the market makers reverse their behavior. With market makers of different patience this strategy gives both a higher profit. LEHRER AND PAUZNER (1999) provide a detailed analysis of these strategies.

When assets trade as complements, additionally to condition (5.73) we need

$$\rho \geq \rho_0' = \frac{\pi_D' - \pi'}{\pi_D' - \pi_E'} \tag{5.76}$$

for the second market maker.

With $\rho_0^* = \max\{\rho_0, \rho_0'\}$ we see that if $\rho \geq \rho_0^*$, i.e. market makers are sufficiently patient, they will not enter their opponents markets. Theorem 6 is valid also in this case with a minimum patience of ρ_0^* instead of ρ_0. Different discount factors obviously could be included into this framework without changing the argument.

Withdrawal of registration as market maker. Suppose that a market maker can withdraw his registration to act as market maker for an asset after T periods and that the other market maker can also only enter after T periods.[79] We use reputation effects to show that in such a situation collusion is still possible.

The condition to satisfy for avoiding market entries, (5.72), becomes

$$\pi_D + \rho\frac{1-\rho^T}{1-\rho}\pi_E + v_1 \leq \frac{1-\rho^{T+1}}{1-\rho}\pi + v_2, \tag{5.77}$$

where v_i denotes the present value of profits after the possible withdrawal for the respective strategies. Obviously we find that upon entering the market the worst reputation can be that the market maker will do so also in the future with certainty, i.e. $\frac{\rho^{T+1}}{1-\rho}\pi_E = v_1$. When not entering, the best reputation is that also in the future he will not enter with certainty, hence $\frac{\rho^{T+1}}{1-\rho}\pi = v_2$.

It is further reasonable to assume that when the market maker has not entered in the past, he is less likely to do so in the future than when he has entered, i.e. he has built up a reputation for not entering (colluding). The less good a market maker's reputation for colluding is, the more likely he defects and receives π_E, hence v_2 decreases as $\pi_E < \pi$. In the same manner it follows that the less good his

[79] Considering intraday transactions, we could define T such that a market maker is only bound for a single trading day as on the NASDAQ.

reputation for defecting is, the more v_1 increases. As $pi_E < \pi$ we will always find that $v_1 \leq v_2$.

These considerations imply $\frac{\rho^{T+1}}{1-\rho}\pi_E \leq v_1 \leq v_2 \leq \frac{\rho^{T+1}}{1-\rho}\pi$.

Constraint (5.77) can be solved for

$$\rho \geq \rho_0'' = \rho_0 + (1-\rho)\frac{v_1 - v_2}{\pi_D - \pi_E} + \rho^{T+1}\frac{\pi - \pi_E}{\pi_D - \pi_E}. \tag{5.78}$$

With the above relations on v_1 and v_2 it can easily be shown that $\rho_0'' > \rho_0$ and collusion is aggravated by requiring market makers to be more patient.

In the special case of $v_1 = v_2$, i.e. the chosen strategies do not influence the future profits after a possible withdrawal, this reduces to

$$\rho \geq \rho_0''' = \rho_0 + \rho^{T+1}\frac{\pi - \pi_E}{\pi_D - \pi_E} > \rho_0''. \tag{5.79}$$

We therewith find that market makers have to be more patient to ensure that they do not enter their opponent's market for assets trading as complements. For assets trading as substitutes both will enter their opponents' markets as before.

As the game does not end after T periods, we can use the results of KREPS ET AL. (1982), MILGROM AND ROBERTS (1982) and KREPS AND WILSON (1982) on reputation effects to ensure market makers not entering the opponent's market, even if they are not that patient. The more important the reputation effect is, i.e. the smaller $v_1 - v_2$ becomes, the less patient market makers can become to ensure that market makers do not enter their opponents' markets as we can see from differentiating (5.78):

$$\frac{\partial \rho_0''}{\partial(v_1 - v_2)} = \frac{1-\rho}{\pi_D - \pi_E} > 0. \tag{5.80}$$

In the limiting case of perfect reputation, $v_1 = \frac{\rho^{T+1}}{1-\rho}\pi_E$ and $v_2 = \frac{\rho^{T+1}}{1-\rho}\pi$, i.e. where the behavior of the past can be taken as given for the future, the same result as in (5.73) emerges with $\rho_0'' = \rho_0$. The absence of reputation effects corresponds to $v_1 = v_2$ and hence condition (5.79). We therewith find that theorem 6 remains valid

with market makers being more patient than without the possibility to withdraw registrations. The minimum patience required to avoid market entries depends on the importance of the reputation effect.

5.5.4 Control structures in larger markets.

In this section we will apply the results derived above to markets with $N > 2$ assets. In such markets, market makers have several options which market(s) to enter and considerations become more complex. This increases the number of possible equilibrium control structures significantly. Furthermore, when already having established a control structure with more than a single asset, portfolio considerations have to be applied to determine the profitability of a market entry.

Suppose again that for every $M \geq 2$ market makers a certain control structure is exogenously given. This control structure can consist of a single asset or any set of assets. These sets are assumed to be different, but not necessarily disjunct.

The initial control structure again is assumed to be an equilibrium, hence profits are equal for all market makers. Furthermore assume that every asset forms part of at least one control structure.

We can now interpret the portfolio of assets a market maker controls as a single, although composed asset and consider the decision to enter the market for an additional asset. In this setting we can use the results from chapter 5.5.2 on the optimal control structure for two assets to determine whether a market entry will be observed.

We distinguish two types of assets for each market maker. One type is the set of assets that trade as substitutes of the portfolio determined by the initial control structure. The other type is the set of assets, trading as complements to this portfolio. Suppose for simplicity that the initial control structure consists only of assets that trade as substitutes.

For assets that are complements to the initial control structure, the result that $\pi > \pi_E$ remains valid also in this more general case. Applying the same intertemporal considerations as in equation (5.72), suggests that complements of his initial control structure will not be controlled by the same market maker if he is sufficiently patient.

Let us now consider the decisions of a market maker to enter markets whose assets are substitutes of his portfolio.

From the results in chapter 4.5.2 it is obvious that the profits from entering a market are larger the closer an asset trades as substitute to the market maker's control structure. From assumption 4 we see that these profits are decreasing the more market makers are registered for an asset.

If the number of registered market makers is sufficiently large, additional profits from entering a market through joint control of assets are outweighed by the loss in market power in the entered market and increased indirect competition . In this case profits will decrease when entering the market, i.e. $\pi > \pi_D$, and the market maker will not enter.

Those market makers whose control structures are close substitutes of an asset will make the largest profits from entering the market, regardless of the number of market makers registered. If the number of registered market makers increases, they would still make profits from entering, whereas market makers whose control structure is less substitutive to this asset will make losses. If a large number of market makers, whose control structure is a closer substitute of the asset, are already registered, they will not enter the market for this asset. Therefore only those market makers will control the asset, whose control structures are the closest substitutes of this asset.

Under which conditions a control structure belongs to the "closest substitutes" of an asset depends on its characteristics. If an asset is very substitutive for many existing control structures this has to be defined in a narrower sense to avoid a too

large number of market makers being registered, such that all encompassed market makers make profits when entering the market.

If a market maker enters a market, the incumbents not necessarily react by entering all markets controlled by the entrant as in the case of two assets. Because market makers differ in their initial control structure and are not allowed to withdraw registrations, their control structures are likely to differ in the characteristics relative to an asset. The symmetry of payoffs in figure 5.12 is no longer reasonable for $N > 2$ assets. Hence it may not be profitable for the incumbent to enter all entrant's markets. He can choose to enter only a few markets the entrant controls or to enter other markets not controlled by the entrant.[80] Therefore equilibrium control structures are likely to be different among market makers and not to encompass all substitutes.

These results are summarized in the following theorem:

Theorem 7. *The equilibrium control structure of a market maker is determined such that it encompasses all those assets for which it belongs to the closest substitutes.*

Securities that trade as complements to the control structure of a market maker will not be controlled by this market maker, provided market makers are sufficiently patient.

In general we find different equilibrium control structures not encompassing all assets trading as substitutes.

It has to be noticed that an asset may be a substitute of one control structure, while being a complement of another, although for both control structures a common third asset has the same characteristics. This shows the complexity of the characteristics

[80] As control structures differ, no implicit collusion across assets can increase profits as has previously been assumed. However it could be a profitable strategy to enter a market and making a loss from this behavior, only to achieve the same control structures and enable implicit collusion between assets to increase profits. However it is unlikely that this strategy will be applied for all assets trading as substitutes, because the large number of market makers registered for each asset would reduce the profits from implicit collusion thus far, that it would not generate larger profits.

an asset can have. This complexity will also be reflected in the equilibrium control structures. Without an exact determination of "closest substitutes" it will not be possible to give precise predictions of the optimal control structure of a market maker. In chapter 5.6.3 we will present a method that might help market makers to identify those assets that are likely to trade as substitutes.

Without the assumption that the initial control structure encompasses only substitutes, we may also find complements to the equilibrium control structure being controlled by a market maker as long as he is not allowed to withdraw his registration.[81]

This analysis suggests that strategic behavior of market makers in determining their control structure prevents them from applying hit-and-run competition to exploit short term profits and therewith enables to execute market power through implicit collusion in the control structure.

5.6 Coordination devices in asset markets

All previous analysis was concerned with finding an equilibrium and determining the best of these equilibria, i.e. the equilibrium with the highest payoff.[82] Although in all cases we found such an optimal equilibrium, it is obvious that for market makers being involved in real markets, these considerations are difficult to make. In the price setting behavior decisions about a new quote have to be made within seconds, so that even with the help of computers a complete analysis cannot be conducted. Furthermore, many other factors affect the payoffs which are not included in the

[81] Such a withdrawal of registration can be prevented by agreements between a market maker and the issuer of an asset that forces a market maker not to leave the market for a specific period of time. Such agreements are required for assets listed on the Neuer Markt of the Frankfurt Stock Exchange. Another reason not to leave the market may be a tradition to control a certain asset.

[82] It may be worth reminding that in the repeated game framework used here, there exist a large number of non-optimal Nash equilibria, e.g. all those prices between the competitive and optimal collusive prices.

models considered here, e.g. adverse selection costs. Whether assets trade as substitutes or complements cannot easily be determined, up to now no general theory has been developed addressing this issue. Finally, relations frequently change over time, which we have not been able to include in our models. All these aspects show that in reality it will be very difficult to choose the optimal equilibrium.

One possibility how to make a decision is to reduce its complexity, e.g. by reducing the strategy space. This reduction may prevent market makers from achieving the highest possible payoff, but still enables him to receive a sufficiently high payoff.[83] As it is of special importance for the sustainability of a collusive equilibrium that the defection of a market maker is detected and punished, these defections should be easily visible. Every market maker using a different, although reduced, strategy space will not facilitate this problem. Therefore only strategy spaces that are equal and reduced for all market makers are useful.

This reduction of the strategy space by all market makers can be interpreted as a coordination game, where market makers choose subspaces of their initial strategy space. If all choose the same subspace(s), they receive a higher payoff due to the easy detection of a defective behavior and hence make higher expected profits as we saw in chapter 5.3.4. Which subspaces are more likely to be chosen, can be analyzed using the theory of focal points as first introduced by SCHELLING (1960).

We will at first briefly present the idea behind focal points and then apply this concept to the price setting behavior as well as market entries of market makers.

5.6.1 The theory of focal points

A problem arising from the existence of multiple Nash equilibria is that with players choosing different equilibrium strategies, the outcome my not be an equilibrium and hence generate nonoptimal payoffs. For this reason players would like to coordinate

[83] That such a behavior is reasonable is subject to the theory of *bounded rationality*. See e.g. KREPS (1998) for an introduction into this topic.

their behavior by choosing strategies that belong to the same Nash equilibrium. Empirical evidence suggests that players are able to coordinate their decisions much better than one would expect if the choice of an equilibrium strategy were made randomly.

To explain this finding, SCHELLING (1960) assumed that certain strategies have attributes, which are not important for the payoff, but attract more attention by players than other strategies. These strategies are called *focal points*. Up to now there exist only few contributions to the theory of focal points, an overview of this literature can be found in JANSSEN (1998). Here we will take a closer look on this approach as presented in SUGDEN (1995), making some minor modifications.

We assume that each player uses a private description of the strategies, called *labels*. Labels are taken from a set Λ^i of all possible labels available to player $i \in \mathcal{P}$. We denote the entity of all labels by $\Lambda = \times_{i\in\mathcal{P}}\Lambda^i$. Define a bijective mapping

$$
\begin{aligned}
L^i \quad : \quad & \mathcal{S}^i \mapsto \Lambda^i \\
& s^i \mapsto L^i(s^i).
\end{aligned}
\tag{5.81}
$$

We can now define a game $\Gamma'(\mathcal{P}, \Lambda, \mathcal{U}, \mathcal{R})$ replacing the initial game $\Gamma(\mathcal{P}, \mathcal{S}, \mathcal{U}, \mathcal{R})$. These games differ only in the way strategies are described, e.g. they are numbered in the initial game and given names in the new game. We assume that this labeling is common knowledge to all players. Suppose now that out of the space of equilibrium strategies, $\mathcal{E}^i \subseteq \Lambda^i$ a subset $\mathcal{A}^i \subseteq \mathcal{E}^i$ is chosen. The space $\mathcal{A} = \times_{i\in\mathcal{P}}\mathcal{A}^i$ we interpret as an advice to the player, or a *recommendation* to choose a strategy from this set.

A recommendation is called *collectively rational* if all players receive lower payoffs from choosing a strategy $L^i(s^i) \in \mathcal{E}^i \backslash \mathcal{A}^i$. We can interpret the set of collectively rational recommendations as *focal points*.

How such focal points are formed is subject to various influences, e.g. culture. Traditions have a very important role in this determination. Strategies that generated

high payoffs in the past, often are assumed to do so in the future, e.g. by forming adaptive expectations.[84] Another aspect may be the attractiveness of "round" numbers on the real line, e.g. integers.

Let us now apply the above model to the framework of repeated games and implicit collusion. As mentioned before, due to the lack of perfect information on all relevant parameters of the other market makers, it is a difficult task to determine whether a market maker is defecting or not. By reducing the set of possible collusive strategies to focal points, it would be easier to detect a deviation. We define focal points, hence select a small set of strategies that are allowed, and interpret the choice of a strategy that is not a focal point as a deviation and punish the market maker accordingly. This punishment induces the player to use a focal point strategy, even if another strategy would be optimal and he intends to collude by choosing this strategy. In this sense he adaptively learns the focal points.[85] CRAWFORD AND HALLER (1990) provide a game theoretic model how players may deviate from optimal strategies to create focal points that will be beneficial to achieve higher payoffs in the future.

The reduction to "allowed" strategies from the set of all equilibrium strategies, restricts the market makers who want to defect. They have to choose a strategy very distinct from their optimal defective strategy to have the chance of being not detected. This reduces his payoffs from defection as well as it increases the risk of being detected due to this large deviation and therefore facilitates sustaining the collusive equilibrium.[86]

[84] According to JANSSEN (1998, p. 154) such a behavior requires players to have bounded rationality.

[85] An example for such a learning process is the market for initial public offerings (IPOs). CHEN AND RITTER (2000) report that for IPOs between USD 20-80 million, the fee charged by underwriters equals 7% for more than 90% of all IPOs. They further report that this fraction has increased constantly over time from less than 25% in 1985. In their investigation they find evidence that this fee is not set competitively and has become a focal point over time to sustain this high level of fees.
In contrast, HANSEN (2000) stresses that underwriters compete in several other dimensions and the 7% rule does not give rise to excess profits. That implicit collusion is not the reason for applying a fee of 7% is supported by the observation that, unlike in the Christie-Schultz debate, the fees have not been changed after the publication of these results, despite antitrust investigations commencing.

[86] A similar situation we faced when introducing discrete prices. There also the strategy space

We will now apply focal points to the price setting behavior of market makers as well as the control structure.

5.6.2 Stock price clustering

With quoted prices set continuously, differences between quotes may only be minor and it is difficult to detect the best available price despite its distribution on the computer screen. The rapidity of decisions will make it likely that a better price is not detected as prices look very alike in the first digits. These costs of determining the best available price as well as the known preference for round numbers gave rise to focal points that restrict quotes to lie on a discrete price grid with a certain tick size. [87] Nowadays all stock exchanges have specified a certain tick size for trading. On US stock exchanges minimum tick sizes of USD $^1/_8$, USD $^1/_{16}$, USD $^1/_{32}$, or USD $^1/_{64}$ are common, with USD $^1/_{16}$ dominating nowadays and USD $^1/_8$ until recently, while European stock exchanges use the decimal system for their minimum tick sizes.

One would expect all price fractions, i.e. 0, $^1/_8$, $^1/_4$, $^3/_8$, $^1/_2$, $^5/_8$, $^3/_4$, $^7/_8$ for a tick size of USD $^1/_8$, to be quoted with equal probability. However, empirical investigations of US assets using tick sizes of USD $^1/_8$ show clear evidence of "round" price fractions to be much more frequently quoted. Investigating a large number of assets, Harris (1991) reports the even eighths quotes 0, $^1/_4$, $^1/_2$, and $^3/_4$ to be quoted in about 60% of the closing prices on December 31, 1987. This pattern is found throughout time, he reports similar findings for selected stocks in 1854.

In their widely recognized study of the most frequently traded NASDAQ stocks, Christie and Schultz (1994) found a much more significant result with 85% of the quotes being on even eighths in 1991. For 70 out of the 100 most actively traded stocks, odd eighth quotes are very seldom found or they were completely absent. If in those cases odd eighth quotes appeared, they were very short-lived and in most

had been restricted to a few well defined strategies and we showed that collusion is facilitated.

[87] In most cases the actual price at which is traded can be different from this price grid.

cases used to widen the spread rather than decreasing it, hence to undercut the current quotes.

We saw how noncompetitive prices and hence spreads can emerge using a repeated game framework with noncooperating market makers. We further saw that without perfect knowledge of all relevant variables, it is difficult to determine whether a market maker has defected or received a combination of parameters requiring him to quote a low collusive price. To better coordinate their behavior, i.e. detect defecting market makers more easily, we can now use the theory of focal points to explain the avoidance of odd eighth quotes.

If market makers concentrate on even eighth quotes, a defective market maker has to undercut his competitors' prices by a quarter rather than an eighth, what reduces his profits from defection significantly and collusion is facilitated as we saw when introducing discrete prices in chapter 5.3.1. If he would quote an odd eighth to limit his losses, the other market makers could easily identify him by this quote as a defector. If a market maker undercuts the current quote by a quarter, it is more likely that his quote is collusive, but based on a different set of variables rather than defective. This reduces the threat of punishment, even if he acts defective.

It is therefore reasonable to assume that quoting only even eighths, or quarters, has become a focal point to maintain collusive pricing more easily. With the same argumentation we could explain the concentration on even sixteenth for assets being quoted in ticks of USD $^1/_{16}$. The public pressure after the publication of the results from Christie and Schultz (1994) to quote competitive prices and the threat of antitrust investigations caused the collusive equilibrium to break down. The costs of maintaining this price regime outweighed its benefits and odd eighths were used more frequently.

After the publication of these findings models have been developed explicitly showing that avoiding odd eighths is generating larger profits to market makers. The two most common models are Kandel and Marx (1997) and Kandel and Marx

(1999a). Empirical evidence given by CHRISTIE AND SCHULTZ (1999) shows that prior to their finding of odd eighths avoidance in cases market makers introduced odd eighths for an asset, the spread increased significantly, while upon withdrawal of odd eighths avoidance the spread decreased. As they could identify no changes in the cost of market making justifying a different spread, it can only be concluded that in cases where market makers were successful in coordinating their quotes on even eighths, they earned excess profits. This finding gives evidence that the use of focal points, i.e. even eighths, can be an appropriate measure to enable implicit collusion. That market makers have chosen even rather than odd eighths can be explained with the already mentioned preference for round numbers.

These empirical results show at first support for the model developed in chapter 5.3. The reduction in the spread after the publication of the results of CHRISTIE AND SCHULTZ (1994) without a significant reduction in the number of market makers registered for each stock, clearly gives evidence that market makers had been able to quote noncompetitive spreads and hence earned excess profits. As in the antitrust investigation following the next years, no evidence on explicit agreements has been found, implicit collusion remains the only explanation for the results. The avoidance of odd eighths quotes furthermore supports the existence of focal points as coordination devices, especially as in combination with the reduction of the spread the use of odd eighths became much more frequent.

GODEK (1996) offers another explanation for the observed stock price clustering on even eighths without assuming implicit collusion between market makers. Define α as the fraction of quotes on even eighths and δ as the difference in the fraction between odd and even eighths:

$$\delta = \alpha - (1 - \alpha) = 2\alpha - 1. \tag{5.82}$$

Even eighths spreads can now occur either by the bid and ask prices to be both on even or both on odd eighths. Define now γ as the fraction of even eighth spreads

relative to all spreads, hence by using (5.82) we get

$$\gamma = \frac{\alpha^2 + (1-\alpha)^2}{\alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)} = 2\alpha^2 - 2\alpha + 1 = 2\alpha^2 - \delta \tag{5.83}$$

The fraction of even eighth spread that are the result of even eighth quotes, θ, is given by

$$\theta = \frac{\alpha^2}{\gamma} = \frac{\delta + \gamma}{2\gamma}, \tag{5.84}$$

which transforms into

$$\delta = 2\gamma\left(\theta - \frac{1}{2}\right). \tag{5.85}$$

Using (5.82) this becomes

$$\alpha = \gamma\left(\theta - \frac{1}{2}\right) + \frac{1}{2}. \tag{5.86}$$

Here the effect that even eighths quotes are more frequently observed is reduced to a high proportion of even eighth spreads, $\gamma > \frac{1}{2}$,[88] and the necessity that these spreads are to a high degree formed by both quotes being on even eighths, $\theta > \frac{1}{2}$, or odd eighths. Using this clustering of even eighths spreads on even eighths quotes due to a preference for even quotes, rather than odd quotes, gives rise to the observed asymmetry in quotes.

GODEK (1996) further points out that he finds significant evidence that those assets exhibiting a large number of even quotes require market makers to quote even eighths spreads to cover their costs. Hence, he concludes, the clustering on even eighth quotes is not necessarily a sign of noncompetitive behavior as in most studies, but the result only of clustering.

In this interpretation even eighths are not used as coordination devices to facilitate implicit collusion. As an additional argument for his interpretation, he shows that after the publication of the results from CHRISTIE AND SCHULTZ (1994) and the

[88] With odd eighth spreads always an odd and an even quote are involved, hence no asymmetry can be found in the quotes.

beginning antitrust lawsuits no significant change to use odd eighths can be observed. Nevertheless, the significant reduction in the spread cannot be explained by clustering.

Despite these two controversial findings and its interpretations, there is a wide agreement among academics that even eighths are used as focal points, although not necessarily as a coordination device to sustain implicit collusion.

5.6.3 Focal points for market entry

In chapter 5.5 we identified the optimal collusive control structure to consist of those assets that trade as the closest substitutes. It has been pointed out that it is very difficult to determine the degree of substitutiveness of assets. Portfolio considerations in multiple asset markets increase these difficulties. We therefore will face severe problems in identifying market makers defecting. For this reason it is useful for them to establish focal points that more easily allow the characterization of assets.

We propose that such focal points are the industrial sector an asset belongs to. Assets belonging to the same industrial sector are more likely to trade as substitutes than assets from different sectors. At first, they exhibit similar return structures as they are influenced by a broad range of common factors with a comparable magnitude. Secondly, portfolio selection of investors frequently forces them to invest into different sectors, such that shifting weights between assets is often associated with shifting shares of sectors, treating all assets of a sector more or less alike. Therefore it is very likely that assets of the same industrial sector trade as substitutes. Informational aspects addressing changes in the allocation of assets within a sector is much less likely to observe.

Of course, also assets from different, but related sectors may trade as substitutes, so that a defection of a market maker can reasonably be assumed only if he controls assets in very distinct sectors. Additionally, historic reasons for controlling certain

assets have to be taken into account, i.e. the initial control structure. Ignoring these historical reasons, we should expect market makers to control assets within a sector and closely related sectors.

Another reason for market makers controlling assets of the same sector can be found in lower adverse selection costs as information on one asset is in most cases also useful for another asset. However, in chapter 4.4 we assumed with BAGEHOT (1971) that market makers are not interested in the fundamental value of the assets they control. For any risks associated with being less well informed they are compensated by the spread. But in most cases market makers are also traders, i.e. they are vertically integrated securities firms. For this reason any empirical results have also to be interpreted on this background.

As no investigation into the control structure can be found in the literature, we conduct an empirical investigation into the control structure of the specialists on the Vienna Stock Exchange (VSE).

On April 1, 1999 the Vienna Stock Exchange (VSE) introduced a new type of market maker, called *specialist*. These specialists are market makers that are obliged to accept larger trades at the stated prices and the maximum spread they are allowed to quote is reduced. Unlike the specialists on the NYSE they are not granted a monopoly in market making, but face competition from "ordinary" market makers. 44 assets were included into this system and for each asset a single specialist has been determined. A total of 10 different specialists have been appointed.

The specialists have been selected according to bids that could be posted by all interested market makers, also if they were not registered as market maker before for this asset. Each interested market maker had to specify the maximum trade size he is willing to accept and the maximum spread he will apply if selected as specialist, where the bids had to meet at least the conditions applied to ordinary market makers. The market maker offering the most favorable conditions has been selected to act as specialist. The specialist is bound by this decision for one year.

Every year new bids from all market makers are collected and the best bidder receives the status of a specialist for the following year. For accepting larger trades and restricting himself to lower maximum spreads, specialists are compensated by paying no transaction fees for trades they conduct as specialist and they receive a share of the transaction fees collected from trades in this asset conducted by other market makers.

To act as market maker in a small market like the VSE, in many cases is not only based on economic considerations, tradition and prestige have an important role. In the market for the 20 most actively traded assets, the Austrian Equity A Market, the five leading banks control on average 17.4 assets, where no bank controls less than 15 assets. But for the 22 less actively traded assets in the Austrian Equity B Market on average only 9 assets are controlled by these banks and no bank controls more than 12 assets.

Although every leading bank has posted a bid to act as specialist for assets she controls as market maker, it can be assumed that the best bids have been posted mainly on economic considerations. For this reason the control structure of the specialists should better reflect economic considerations than the control structure of the market makers.

Only five of the ten specialists on the VSE control more than two assets: Bank Austria AG (BA), Creditanstalt AG (CA), Centro Internationale Handelsbank AG (Centro), Erste Bank der österreichischen Sparkassen AG (Erste) and Raiffeisen Zentralbank Österreich AG (RZB). These five banks control 38 of the 44 assets that are traded using specialists. As it is of interest to investigate the control structure, only these five specialists are included in the investigation, excluding the five other specialists.

It is assumed that assets of firms in the same sector trade as close substitutes, whereas assets of firms in different sectors trade as less close substitutes or complements. The sectors used in the analysis are those the companies are assigned to

	BA	CA	Centro	Erste	RZB
Machines and metal processing	3	2	1	1	3
Food	0	2	1	0	1
Finance	4	0	0	2	1
Energy	0	0	3	0	0
Construction and real estate	2	0	2	0	1
Other	2	4	0	1	2

Table 5.3: Control structure of VSE specialists

by the VSE, where some sectors have been aggregated in order to have a sufficient number of assets in each sector. The used sectors are: (1) Machines and metal processing, (2) Food, (3) Finance, (4) Energy, (5) Construction and real estate and (6) Other. The control structure of the five specialists is shown in table 5.3, the entries are the number of assets a specialist controls in the respective sectors.

If the theory of the optimal control structure from chapter 5.5 is correct, it should be observed that specialists concentrate on one or two sectors, i.e. the control structure should not be stochastically independent. Computing the statistics for classified data to test for stochastic independence, the value turns out to be $\chi^2 = 31.56$. The test statistic is χ^2-distributed with 20 degrees of freedom, hence stochastic independence is rejected at the 95%-level.[89]

As the sector "Machines and metal processing" is very heterogeneous, it only contributes marginally to this result, whereas the relatively homogeneous sectors Finance and Energy contribute most. When aggregating sectors (1), (2) and (5) to a sector called "Industry", the result does not change. The result also holds if the sector "Other" is eliminated from the investigation in the original investigation as well as with the aggregated sector "Industry". In both cases stochastic independence of the specialist control structure is rejected at the 95%-level.

[89] The test statistics for the independence of classified data is called χ^2 and χ^2 distributed with $(n-1) \times (m-1)$ degrees of freedom, where n denotes the number of columns and m the number of rows of the data. See BAMBERG AND BAUR (1993, pp. 202 f.).

These results seem to support the theory, although they have to be interpreted with care. The small sample size of only 38 assets and five specialists is too small to rule out other factors influencing the control structure.[90] Investigating larger markets, like the NASDAQ with more than 5,000 assets and 500 market makers, may bring further insights. It may also be useful to analyze more detailed whether assets trade as substitutes or complements as this has been done here.

Interestingly, when analyzing the structure of the posted bids to act as specialist as well as the market maker structure of the VSE, stochastic independence cannot be rejected at any reasonable level of significance. This supports the above remark that to act as market maker is also driven by other than only economic considerations. The same holds for posting a bid to act as specialist. It can be assumed that in most cases a bid has been posted that did not significantly improve the accepted trade size and the maximum spread. Bids were posted hoping that no other market maker offers better conditions and that they can make additional profits from receiving the benefits of a specialist without reducing the profits from market making substantially.

We can conclude this section by stating that through focal points empirical support can be found for implicit collusion in the price setting behavior as well as in the control structure.

5.7 Summary

In this chapter we investigated strategic behavior of market makers in dealer markets. Unlike in the traditional market microstructure theories we were not concerned with the costs of market making and their implications for the price setting, but

[90] After the reallocation of specialist positions in 2000, we observe a concentration of these positions in the hands of Bank Austria AG, Creditanstalt AG and Centro Internationale Handelsbank AG, which control 30 out of the now 42 assets trading under the specialist system. This development caused that it is no longer possible to reject stochastic independence of the control structure. However, we observe that while Centro Internationale Handelsbank AG controls 14 assets in all sectors, Bank Austria AG and Creditanstalt AG concentrate on different sectors as predicted by the theory on optimal control structures.

concentrated on the possibility of market makers to reduce competition through strategic behavior. This reduced competition allows them to execute market power and therewith receive excess profits at the costs of investors.

We investigated two possibilities how market makers can earn excess profits: by colluding to quote wider spreads and by colluding to use a control structure reducing indirect competition between assets and avoiding hit-and-run competition.

It was found that collusion between market makers to quote wider than competitive spreads is sustainable under fairly general and realistic conditions, even for actively traded assets with a large number of market makers. We furthermore showed that the properties of spreads under implicit collusion are not sufficiently different from those under competition, such that we cannot expect to find any direct evidence for or against the presence of implicit collusion from market data.

Reducing indirect competition between assets through an appropriate control structure was also found to be sustainable. Here the results should be empirically testable by investigating the control structure of market makers. However, such an investigation will be aggravated by the difficulties in determining the properties of assets that are needed to interpret the results.

All results on implicit collusion were derived by assuming market makers to behave strategically, i.e. take into account the consequences their actions have on future profits. If they are sufficiently patient, i.e. discount future profits not too much, they will abstain from exploiting a possible current profit at the cost of reducing future profits. In this sense a strategic market maker behaves less competitive today to receive larger profits in the future.

As direct evidence on such behavior is difficult or virtually impossible to find, we took into account the complexity of decisions that have to be made within a very short period of time. It would be unreasonable to assume market makers conducting such sophisticated calculations as in the models developed in this chapter. They are more likely to reduce the complexity of their decisions, e.g. through the reduction of the

possible alternatives they have to consider. We found evidence for the existence of such focal points, e.g. the odd eighths avoidance by market makers on the NASDAQ which gave rise to the Christie-Schultz debate. Through the existence of such focal points, we have presented indirect evidence on the existence of implicit collusion among market makers.[91]

All these aspects show that strategic behavior of market makers can have significant influence on the trading costs for investors and should therefore no longer be ignored in market microstructure analysis. The models developed above may serve as a starting point for a more sophisticated analysis of strategic behavior, including strategic behavior of investors. Such a more general analysis could include the (strategic) submission of limit orders, strategic behavior of informed investors or competition from other trading facilities, like ECNs.

Throughout this chapter we assumed the number of market makers for an asset or for the entire market to be exogenously given. By assuming a certain fixed cost to become a market maker, we could easily derive the equilibrium number of market makers in analogy to the model of Grossman and Miller (1988) presented in chapter 4.1.1. However, such a situation is not very realistic. The access to the market is very restricted, with new market makers being required to pass very strict entry barriers, even on the NASDAQ. In several other markets no additional market makers are admitted. Furthermore, nearly all leading securities companies have already access to the most important markets, such that additional market makers are difficult to find. Even if such new potential market makers were found, we could easily expand our model to show that implicit collusion arising from multimarket contacts in other areas than market making, e.g. investment banking or investment consulting, will prevent them from becoming market makers.

[91] In chapter 6.1 we will provide more detailed evidence on implicit collusion in the price setting behavior among NASDAQ market makers.

Chapter 6

Policy measures to reduce the effects of implicit collusion

As has been pointed out in the introduction, low transaction costs are a major concern of stock exchanges. It has also been mentioned that the spread is a main source of transaction costs in dealer markets.

We saw in the previous chapter that even noncooperating market makers can implicitly collude in the price setting behavior as well as the determination of their control structure to quote spreads above competitive levels. Implicit collusion was further shown to be very difficult to detect, direct measures against this behavior can therefore not be implemented. In this chapter we will consider several policy measures that can reduce the ability of implicit collusion among market makers. Empirical investigations as well as theoretical contributions will be used to evaluate the effectiveness of these policy measures.

Most attention has been given to the Christie-Schultz debate commencing in 1994 and the revision of order handling rules on the NASDAQ in 1997, enabling limit orders to compete directly with the quotes set by market makers. We will therefore start our discussion of policy measures with a detailed review of this literature. Afterwards the focus will be laid on other aspects that have been given only very limited attention in the recent discussion: the access of new market makers, market transparency, the question of time priority, restrictions on the size of the spread, anonymity in trading, and tick sizes.

6.1 Competition from limit orders

The models discussed in chapter 5 assumed throughout that all trades have to be conducted using market makers. As has been pointed out in chapter 4.5.1, limit orders also provide liquidity and for investors quotes set by a market maker and limit orders are equivalent. We will therefore in this section investigate whether it is beneficial to allow for competition between quotes set by market makers and limit orders. As most empirical contributions address the NASDAQ, we have to base our conclusions mainly on these findings.

Until 1994 limit orders have been interpreted by the NASDAQ as offers to trade with a market maker rather than with other orders in the market. Market makers only had to execute limit orders if their quotes enabled them to do so, but they were allowed to trade with other investors at less favorable prices than the limit order price. In the aftermath of the CHRISTIE AND SCHULTZ (1994) findings on odd eighth avoidance[1] this rule was changed such that market makers having received a limit order must not trade with other investors at less favorable prices than the limit order price. Other market makers were not restricted in their trading and the market maker having received the limit order was allowed to trade for his own account at the same price.

In two sets of each 50 assets from the NMS on January 20, 1997 and February 10, 1997 limit orders had to be published by the market maker receiving limit orders as if they were there own quotes.[2] A market maker was no longer allowed to trade for his own account at a less favorable or the same price with other investors. Since the introduction of this rule change, limit orders can directly compete with quotes set by market makers.[3]

[1] See chapter 1.1 for a brief review of the Christie-Schultz debate.

[2] Instead of publishing the quote they were also allowed to execute the limit order immediately on their own accounts. In the mean time this rule is applied to all assets trading on the NASDAQ.

[3] At the same dates as mentioned above, prices offered by ECNs also had to be incorporated into the quote system, adding an additional source of competition for the quotes set by market makers. Empirical investigations of the effects arising from this change of the order handling

There are generally two groups of investigations. One group of studies compares trading costs between stocks listed on the NASDAQ and stocks listed on another stock exchange. They either identify two groups of stocks with similar characteristics and compare the spread between these two groups, investigate stocks dually listed on stock exchanges with different market structures, or they investigate the spread of stocks before and after they change their listing from one stock exchange to another. The other group of studies compares the spread of the same stocks listed on the NASDAQ before and after the order handling rules have been changed.

BARCLAY (1997) investigates a large number of stocks changing their listing from the NASDAQ to the NYSE or AMEX prior to the publication of the results by CHRISTIE AND SCHULTZ (1994). He observes that those stocks trading on the NASDAQ mostly on even eighths, experience a much larger spread on the NASDAQ than those traded also on odd eighths. When changing their listing to the NYSE or AMEX this difference vanishes, all stocks quote on even and odd eighths and the spread is approximately the same for both groups. The reason for this change is the much larger reduction in the spread of stocks formerly quoting mostly on even eighths. This result suggests that the competition from limit orders reduced the spread.[4] Similar findings are reported by HEIDLE AND HUANG (1999). Unfortunately, no investigation exists using data from stocks having changed their listing after the publication of the results by CHRISTIE AND SCHULTZ (1994), such that it is not possible to determine whether the effect is still observable in the current environment.

In 1987 the Chicago Stock Exchange (CHX) introduced trading in NASDAQ stocks. In contrast to the NASDAQ, the CHX is a specialist market like the NYSE or AMEX, i.e. a single market maker (specialist) competes with limit orders for trading volume. This dual trading of stocks in two different market structures allows to

rules are aggravated by the other market reforms that have become effective on June 2, 1997: the reduction of the tick size from USD $^1/_8$ to USD $^1/_{16}$ and the reduction in the minimum order size a market maker has to accept.

[4] On the NYSE and the AMEX limit orders could always compete directly with quotes set by monopolistic market makers, called specialists.

compare trading costs directly. Although the market share of the CHX was less than .8% in 1993 and reduced further to .18% in 1996, i.e. liquidity was much lower on the CHX than on the NASDAQ, VAN NESS ET AL. (1999a) and VAN NESS ET AL. (1999b) report that the spreads on the CHX are smaller than on the NASDAQ. This results supports the hypothesis that competition from limit orders reduces trading costs. These findings are in line with many other empirical investigations comparing trading costs in dealer markets and markets dominated by limit orders, several of these studies are mentioned in VAN NESS ET AL. (1999a) and VAN NESS ET AL. (1999b). VAN NESS ET AL. (1999b) further report that on the CHX even and odd-eighths are equally distributed, while they also find a clustering on even-eighths in NASDAQ quotes.

Empirical results, e.g. the recent study by CHUNG ET AL. (1999a), besides theoretical considerations like DEMSETZ (1997), stress the importance of the competition between limit orders and specialists on the NYSE. If the lack of competition from limit orders is the reason for the spreads to be larger on the NASDAQ compared to similar assets listed on the NYSE, as suggested by the above mentioned findings on stock listings, we should observe a substantial decline in the spread after limit orders were allowed to compete directly with the quotes set by market makers. Comparing the spread before and after the publication of the findings in CHRISTIE AND SCHULTZ (1994), CHRISTIE ET AL. (1994) and SCHULTZ (2000), show a decrease in the market spread[5] of about 40% and odd eighths are much more frequently used than before. Nevertheless, spreads on the NASDAQ remain larger than on the NYSE or AMEX, but the difference has decreased substantially.

As at this time no change in the order handling rule had come into effect, this reduction cannot be attributed to the competition from limit orders. The regulatory, legal and public pressure on the market makers can be assumed to have caused the collusive agreement to break down at least partially. Most importantly, market makers lost their ability to coordinate their quotes on even eighths.

[5] The market spread, also called inside spread, is the difference between the best available bid and ask prices.

The change in the order handling rules in 1997 has been shown in NASD, INC. (1997), BARCLAY ET AL. (1999), BESSEMBINDER (1999) and WESTON (2000) to reduce the market spread further by approximately 30%, whereas individual market maker spreads decreased only by 11%. These results give evidence that the dispersion of quotes has increased due to the inclusion of limit orders into the quotes of market makers. But as it is no longer possible to distinguish between the quotes set by a market maker and a limit order, the fraction due to smaller spreads in the quotes of market makers cannot be inferred. For investors the spread decreased significantly with the market reform.

CHUNG ET AL. (1999b) investigate the spread on the NASDAQ after the reforms of 1997 and find that compared to the NYSE, there still exists a clustering on even-sixteenth, but not any longer on even-eighth.[6] They further show that this clustering cannot be explained by characteristics of the assets or the trading process, suggesting that there still exists a smaller, but persistent effect arising from implicit collusion and the coordination of quotes.

DORAN ET AL. (1995) find that NASDAQ market makers adjust their quotes much less frequently than the quotes on the NYSE. These frictions, attributed to the larger amount of limit orders on the NYSE, account also for a part of the larger spreads. If it takes more time for market makers to revise their quotes, we will more frequently observe trades being executed at less favorable prices and spreads are more likely to be larger.

That competition from limit orders only reduces the effects from implicit collusion but does not abandon this behavior, maybe due to the finding that individual as well as institutional investors prefer to submit limit orders on even quotes as found by COONEY ET AL. (1999). This preference for submitting limit orders at even quotes reduces the competitive force arising from them and allows market makers to sustain their arrangement.

[6] As previously mentioned, on June 2, 1997 the NASDAQ changed the minimum tick size from USD $^1/_8$ to $^1/_{16}$.

In contrast, He and Wu (1999) attribute the differences remaining after the reform of the NASDAQ to higher adverse selection costs resulting from more informed investors in the market and conclude that the competition from limit orders on the NASDAQ has been sufficient to eliminate effects arising from implicit collusion.

Kleidon and Willig (1995) question the existence of implicit collusion in the results of Christie and Schultz (1994) at all. They argue that the use of odd-eighths quotes not necessarily implies implicit collusion. The use of odd-eighth quotes in the past has a statistically significant impact on the current use of odd-eighth quotes. This argumentation is in line with Godek (1996), who showed that if even eighths spreads are more frequently found, we can expect to find more frequently even eighth quotes than odd eighths quotes.[7] More interestingly, they stress that multiple competitive equilibria, e.g. one equilibrium with a large spread and a high depth of the market and another equilibrium with a low spread and a small depth, may exist. The regulatory pressure after the publication of the findings by Christie and Schultz (1994) caused the equilibrium to change. Hence we can find less even-eighths, smaller spreads and lower depths even if market makers did not implicitly collude before.

Similar rule changes on the London Stock Exchange in October 1997 also allowing limit orders to compete more directly with quotes set by market makers, have also reduced the spread significantly as reported by Naik and Yadav (1999). These findings suggest that the results obtained from the NASDAQ are generally valid in dealer markets.

Biais et al. (1995) show that only with direct competition between market makers and limit orders, the competitive equilibrium is the only equilibrium. If limit orders do not compete with quotes set by market orders, at least one noncompetitive equilibrium exists.

We can summarize the findings of the above studies briefly as follows:

[7] See chapter 5.6.2 for a derivation of this result.

- spreads in dealer markets with direct competition from limit orders are smaller,
- the spreads in NASDAQ stocks reduced with the publication of the results on odd eighths avoidance prior to the change of the order handling rules,
- market spreads reduced further with limit orders directly competing the quotes set by market makers.

In the light of these results, despite some objections like the possibility of multiple equilibria, it seems to be obvious that the competition from limit orders aggravates to sustain implicit collusion, reduces the spread and therewith trading costs for investors. This gives rise to our first policy implication:

Policy implication 1. *Limit orders should directly compete with quotes set by market makers.*

6.2 Market access

The analysis in chapter 5.3 showed that market makers have to become more patient the smaller their share of the order flow is and the incentive constraint for quoting collusive prices becomes more binding, what reduces the spread for larger liquidity events. As with a larger number of market makers the share of the order flow for every single market maker weakly decreases, the effect arising from implicit collusion is reduced and the average spread becomes smaller. For this reason exchanges should be interested in a large number of market makers being registered for each asset.

Investigating NASDAQ stocks, Wahal (1997) and Klock and McCormick (1999) show that the average spread of an asset reduces if the number of market makers increases. Similar results are obtained from theoretical models, as in Nelling and Goldstein (1999), and experimental settings as reported by Krahnen and Weber (1997) and Ackert and Church (1999). However, as we have shown in chapter 5.5.2, market makers also implicitly collude to achieve a control structure

that maximizes their profits through optimal execution of market power. These considerations restrict the number of market makers for an asset even with free entry of market makers to register for a specific asset, but a limited number of market makers admitted to the entire market.

One could now argue that admittance to the market should be completely free. But as market makers not only have the obligation to quote both, bid and ask prices permanently, but market quality depends on their behavior, this option would not be optimal. Inexperienced and uneducated market makers are likely to make large losses on their activity, such that the possibility of bankruptcy increases and hence the counterpart risk, increasing transaction costs for investors. If the exchange guarantees an order to be fulfilled at the stated conditions, it has to charge a fee for this guarantee, also increasing transaction costs. Therefore it seems to be more useful to restrict access to the market. But with limit orders competing with quotes as described in chapter 6.1, those market participants not given the status of a market maker can compete their quotes by placing limit orders.

Another aspect showing that a too large number of market makers does not result in lower spreads, is presented by LAUX (1995).[8] He also infers the number of market makers to decrease excess profits. But on the other hand, for any given number of transactions in a certain period of time, the more market makers are present, the fewer trades each market maker receives. If a market maker has an inventory position, this means that he has to wait longer until having the chance of receiving an offsetting order. With the fundamental value of the asset changing over time, this results in a higher risk and therewith as shown in chapter 4.3.1, his inventory costs increase. Furthermore (fixed) order processing costs increase per traded share.

Additionally, informed investors could trade more frequently by submitting several smaller orders to the market, which are executed subsequently by the market makers before they can update their quotes. Hence the share of informed order

[8] A similar result has already been derived by GARBADE AND SILBER (1979) in a different framework.

flow increases, and according to the results in chapter 4.4.1 adverse selection costs increase.

therewith we have identified two competing forces: excess profits decrease and the costs of market making increase in the number of market makers. The optimal number of market makers giving the smallest spread is therefore finite.

Outside competition imposes a similar threat to inventory costs of market makers, but to a much smaller extent, because limit order traders have not the obligation to quote prices permanently like market makers. Furthermore, higher costs of limit order trades compared to the costs of market makers, e.g. broker fees, limit the extent of outside competition. For this reason outside competition can be expected to increase the costs of market making to a smaller extent than additional market makers.

We can conclude our findings with the following policy implication:

Policy implication 2. *The number of market makers admitted to the market should be limited.*

6.3 Market transparency

Throughout all models we assumed strict price priority, or equivalently that price information is available to all market participants for free. Such an assumption is reasonable for markets like the NASDAQ where prices are published and disseminated electronically and small orders are, if not preferenced, automatically routed to the market maker quoting the best available price. But in other markets, like telephone markets, prices are not always published. Therewith a broker having the obligation to execute an order at the best available price has to search for this price, what imposes costs on him. These costs he will charge to the investors, whose transaction costs are increasing.

The presence of search costs will prevent a broker from searching the entire market, but only a fraction. Hence he will not in all cases find the best available price and execute orders at a less favorable price.[9] This causes the demand for the market makers not quoting the best available price not to reduce to zero and a defecting market maker will not serve the entire market. This situation is comparable to preferencing arrangements, where also market makers not quoting the best available price receive the preferenced share of the order flow. According to the results in chapter 5.4 collusion is facilitated in intransparent markets.

NILSSON (1999) investigates the effect of price publication, i.e. reduced search costs, on the ability to sustain the collusive equilibrium. His evidence is mixed by finding two competing effects. Price publication enables to apply strict price priority, hence incentives to deviate are increased. On the other hand the punishment of a defective behavior becomes more difficult as competitive pricing has to be applied if a punishing market maker wants to have the chance of receiving at least a small fraction of the order flow. Depending on parameter constellations, search costs, patience and share of the order flow, collusion is either facilitated or aggravated.

In the case that search costs are zero with price publication, as can reasonably be assumed with automated order routing and electronic publication of the price, it is shown that collusion is aggravated. FLOOD ET AL. (1999) confirm this result in an experimental setting. They find that the spread is lower in markets that publish their quotes.

Besides the publication of quotes, market transparency also includes information on trades conducted. While the NASDAQ publishes information on trades instantly, many of the new emerging ECNs do not. In these markets, e.g. Instinet or Island, we find that spreads are in most cases lower than for the same asset trading on the NASDAQ. BLOOMFIELD AND O'HARA (1999) and BLOOMFIELD AND O'HARA (2000) investigate an experimental asset market with adverse selection costs im-

[9] In chapter 4.1 we have shown that these search costs make the existence of a central match or market maker beneficial for investors.

posed on market makers to show that in markets without information on trades, spreads are smaller. This is due to the fact that market makers can use their private information on the order flow to reduce adverse selection costs, while in markets reporting information on trades this information can also be used from informed and uninformed investors to change their believes, such that the market maker has no informational advantage from observing the order flow. However, this intransparency comes at a price; markets are informationally less efficient, which imposes additional costs on investors.

In contrast, PAGANO AND ROËLL (1996) provide a theoretical model showing that on average trading costs for investors are lower in transparent markets. However, for certain order sizes trading costs for investors can be increasing.

Whether information on trades should be published or not depends on the goals of the market structure. If the aim is to create a market structure allowing for high informational efficiency, trades should be reported. Whereas if the aim is to have a market with low trading costs, information on trades should not be published.

Despite these impacts on informational efficiency and trading costs for investors, it has to be noticed that in cases where neither quotes nor transaction prices are published, defection from implicit collusion is very difficult to detect. Market makers can only infer from their individual order flow whether another market maker is likely to defect. As has been pointed out in chapter 5.3.4, this reduces the prices for medium liquidity events and increases prices for small and high liquidity events. HUCK ET AL. (2000) show that in markets where market makers have more information about their competitors' quotes and costs, implicit collusion is aggravated and the outcomes are more competitive.

Summarizing this discussion we can state two competing effects: on the one hand market transparency increases the efficiency of markets, whereas adverse selection costs increase and the publication of quotes aggravates implicit collusion. On the other hand, if neither quotes nor trades are published, the detection of implicit

collusion becomes more difficult, increasing trading costs for small and large liquidity events and decreasing spreads for medium liquidity events.

Balancing these aspects we come up with the following policy implication:

Policy implication 3. *Quotes should be published by the exchange and be available for free.*

If the main goal of the market is informational efficiency also information on trades should be published, if it is the aim to offer an environment for trading at low transaction costs, trade information should not be published.

6.4 Time Priority

As has been pointed out previously, quotes set by market makers are not subject to time priority, i.e. the execution does not depend on the time a quote has been posted.

With time priority the profits arising from defection would be larger. If a market maker reduces the quote at the ask or increases the quote at the bid, it is very likely that other market makers will follow his step and quote the same price. With time priority he would receive the entire order flow upon his improvement of the price and therewith make a higher profit than without time priority, where he receives only a fraction of the order flow.

therewith incentives for price competition would be increased if market makers were also subject to time priority. Cordella and Foucault (1999), and Simaan et al. (2000) provide theoretical models to derive this conclusion stated above intuitively.

On July 29, 1996 the Toronto Stock Exchange changed its rule from time priority to a more complex priority rule, only partially having included time priority and adding size priority. An empirical investigation by Angel and Weaver (1998)

showed that upon this rule change, price competition decreased and size competition increased. They found the average spread to increase significantly, which confirms the results of CORDELLA AND FOUCAULT (1999) and SIMAAN ET AL. (2000).

These considerations give rise to the following policy implication:

Policy implication 4. *Time priority for quotes set by market makers increases price competition and therewith reduces the effects arising from implicit collusion.*

6.5 Restrictions on the spread

A natural way to limit the extent of implicit collusion seems to be through imposing restrictions on the size of the spread. Such a restriction could either be defined as a fixed maximum spread or a maximum spread to depend on market conditions.[10] But these restrictions also have adverse effects on the spread.

With a fixed spread this restriction has to be sufficiently large such that in times of market stress it still compensates market makers for their increased costs. Therefore under normal trading conditions the restriction on the spread does not come into effect and cannot reduce implicit collusion. If the maximum spread is too small to compensate market makers in extraordinary situations, they will quote larger spreads in normal times and subsidize these extraordinary events. On average we can therefore expect to find larger spreads. Furthermore, spreads cannot be set according to the costs of market making, resulting in an inefficient allocation of resources.

With spread restrictions defined relative to the spread quoted by other market makers, like on the NASDAQ, we cannot expect spreads to decrease at all. With all market makers colluding, the quoted spread would not be required to be reduced.

It is more likely that spread restrictions serve as focal points and therewith larger spreads will be applied on average.

[10] The NASDAQ requires market makers to quote an monthly average spread of no more than 150% of the average quoted spread of all market makers.

Policy implication 5. *Maximum spreads are not an appropriate measure to reduce the effects arising from implicit collusion.*

6.6 Anonymity in markets

In chapter 5.4 we showed that order preferencing arrangements facilitate implicit collusion. Therefore measures should be taken that these arrangements are not enforceable. RICKER (1998) points out that preferencing arrangements prevent market makers from competing on price as they receive a share of the order flow also if they do not quote the best available price, hence they still can make a profit.

In anonymous markets where only the quotes of market makers are published, but their identity is not given and the orders are routed to the market maker with the best available price electronically, no such order preferencing arrangement can survive.

From this point of view the "electronic" order book seems to be desirable.[11] On the other hand does such an anonymous market not allow to discriminate certain groups of investors. It is common in markets with order preferencing arrangements that small orders are executed at more favorable prices than the quoted prices, i.e. they are subject to quote improvements, as in most cases they are liquidity induced and do not impose significant adverse selection costs on the market maker.[12]

In an anonymous market with electronic order routing and hence anonymity of trading we find that small orders face higher costs than before as they cannot be

[11] GLOSTEN (1994) provides a model showing that such an electronic order book is optimal and no other market form can successfully compete.

[12] On the NASDAQ it has been a concern that informed investors can submit a larger number of small orders instead of a single large order. These orders, which impose adverse selection costs on market makers, are then executed using the SOES system originally designed for small, and hence uninformative orders. As these informed investors exploit the SOES system, they are also called *SOES bandits*. With such a behavior the advantages of preferencing arrangements are reduced. KANDEL AND MARX (1999b) provide a model how market makers can avoid too much losses from SOES bandits. They show that odd eighths avoidance, i.e. a larger implicit tick size can be an appropriate measure.

treated separately, while larger orders may face lower transaction costs. THEISSEN (1999) investigates the German Stock Market and finds that in the anonymous trading of the XETRA system, adverse selection costs are generally higher, what confirms our claim.

However, we suggest that the reduction of implicit collusion can outweigh the effects arising from higher transaction costs, especially if no other measures considered thus far are undertaken. With competition from limit orders, implicit collusion may be reduced thus far that an increase in transaction costs turns out for the average investor.[13]

We therefore propose the following policies:

Policy implication 6. *If no other measures are applied to reduce the effect of implicit collusion the introduction of an anonymous market with electronic order routing can help to reduce the effects arising from implicit collusion.*

With other measures to limit the effect of implicit collusion on transaction costs, order preference arrangements may reduce trading costs for small orders and are therefore beneficial.

6.7 Tick sizes

We showed in chapter 5.3.2 that implicit collusion is facilitated the larger the tick size is. BERNHARDT AND HUGHSON (1996) also claim that discrete pricing reduces price competition and allows market makers to quote noncompetitive prices. In the

[13] To the author's best knowledge no empirical investigation into the relation of preference arrangements and spreads for the NASDAQ exists. The only investigation into this topic addresses the London Stock Exchange, for which thus far no evidence for implicit collusion has been reported. HANSCH ET AL. (1999) find evidence that preferenced trades are subject to higher spreads, which they attribute to the lower search costs for brokers to find the best available price. They cannot find any support for a positive relation between the degree of preferencing and the quoted spread. Therefore this finding cannot be used to support the hypothesis that preferencing facilitates implicit collusion and for this reason should be prevented.

light of these results it seems to be desirable to reduce the tick size as much as possible. But there arise several objections from such a proposition.

First of all, empirical investigations suggest that upon reducing the minimum tick size from USD $^1/_8$ to USD $^1/_{16}$ on the NYSE as well as on the NASDAQ, the spread declined significantly by approximately 25% as reported by GOLDSTEIN AND KAVAJECZ (2000), RICKER (1999) and NASD, INC. (1997). Further evidence from the Toronto Stock Exchange, the AMEX and the London Stock Exchange, as reported in RICKER (1999) and HUANG AND STOLL (1999), as well as projections from the NYSE before they changed to a tick size of USD $^1/_{16}$ in HARRIS (1994), gives similar results. On the other hand, the same studies give evidence that the depth[14] of the market reduces. Hence larger orders can only be traded at less favorable prices and only small orders benefited from the reduction of the tick size. Empirically, JONES AND LIPSON (2000) show that trading costs on the NYSE increased for large orders after the tick size has been reduced in June 1997.

In their theoretical contribution, CORDELLA AND FOUCAULT (1999) show that larger tick sizes do not necessarily give rise to less competitive results. The models presented in chapter 5 assumed that all market makers post their quotes in a simultaneous move, while CORDELLA AND FOUCAULT (1999) assume these prices to be set sequentially by market makers. They allow them further to revise their quotes if they are undercut by a competitor.

In this setting the quotes converge to their equilibrium level over time. Typically, prices are undercut by a single tick size. If we assume that it is costly for market makers to observe their competitors' quotes, they will not continuously be prepared to react to the update of a competitor's quote. The more steps have to be taken for a price to adjust to its equilibrium, i.e. the smaller the tick size is, the longer this process will take. An order arriving on the market is for this reason more likely to be executed while the adjustment process is not completed and hence receive a less

[14] The depth measures the order size a market maker is willing to accept at the stated price, or in cases of limit orders, the size of the best limit order.

favorite price.

This gives rise to two competing effects. On the one hand, smaller tick sizes decrease the effects arising from implicit collusion and therewith ensure a reduced spread in equilibrium. On the other hand, the longer time to adjust prices to the equilibrium makes it more likely that orders execute at less favorable prices. For large tick sizes the first effect dominates, while for small tick sizes the latter is more important. CORDELLA AND FOUCAULT (1999) show that the optimal tick size is not the smallest possible, but an intermediate. We may use this result also to point out that focal points in the price setting may be favorable to investors in some instances.

The current debate in the United States about decimalization, i.e. changing the ticks from multiples of USD $^1/_{16}$ to the decimal system, in most cases is associated with reducing the tick size to USD .01.[15] An argument that a decimal system is favorable to weird fractions would be that it is more convenient to all market participants as they are used to the decimal system in ordinary life. Furthermore, nearly all stock exchanges over the world apply a decimal system nowadays. However, in light of the discussion above it may be worth considering whether a tick size of USD .01 is suitable or a larger tick size should be implemented.

If market makers want to collude implicitly, the change to a decimal pricing will not prevent them from doing so. To coordinate their behavior, they will find new focal points, probably USD .05 or USD .10. Only in a transition period we can expect that implicit collusion is aggravated until the new focal points are found.

We may summarize this discussion in a final policy implication:

Policy implication 7. *Reducing the tick size to its minimum in general does not reduce trading costs.*

[15] RICKER (1999) gives an overview of this debate. Further comments can be found in WHITCOMB (1997) and RICKER (1997). In a release published on June 13, 2000, the SEC requires US stock exchanges and the NASDAQ to introduce a decimal system no later than April 9, 2000, see SECURITIES AND EXCHANGE COMMISSION (2000a).

6.8 Summary

This chapter considered several policy measures to aggravate implicit collusion and limit its effects on trading costs. We showed that the most appropriate measures are to allow limit orders to compete the quotes set by market makers, introducing time priority for the quotes of market makers and if no other measures are taken, anonymity of trading.

Other measures such as imposing a maximum spread, admitting a larger number of market makers or a reduction of the tick size, show also adverse effects on the spread and therewith transaction costs of investors.

A long-lived discussion in this field is the superiority of competing dealer markets (like the NASDAQ) or auction based specialist markets (like the NYSE). Empirical investigations show lower spreads on the NYSE than on the NASDAQ when controlling for all differences in the characteristics of assets. The competition of limit order trading reduces the spread further than competition between market makers. On the other hand, competing dealer markets have the advantage of faster information revelation as reported by Heidle and Huang (1999), such that adverse selection costs for investors are smaller. These divergent aspects illustrate the claim made in chapter 2.3 that every market form has its advantages and disadvantages. To determine the optimal market form, all aspects have to be balanced accordingly.

When having evaluated the policy measures, we mainly focused on the spread, i.e. transaction costs. Aspects like informational efficiency or liquidity only received limited attention as the focus of this work is on the spread. In a complete analysis of policy measures, however, these aspects have to be considered in much more detail. Angel (1997) provides a detailed analysis of optimal trading rules for stocks of small companies. He shows that the characteristics of the assets as well as of the investors are very important in these considerations.

Concluding remarks

In this work we investigated strategic behavior of competing market makers in dealer markets and laid a special emphasis on its implications for the quoted spread.

In recent years increased competition between exchanges forced every market to offer a market structure which allows investors to trade at low costs. Such a market structure enables to attract a high trading volume and new listings. As has been pointed out in chapter 1, the spread quoted by market makers forms an important part of these trading costs, although not the only part. By reducing the spread, enhancing informational efficiency and continuously improving the technology of the exchange to guarantee faster access to the market at low costs, exchanges want get an advantage over their competitors.

In this work we concentrated on the spread as one of the most important source of trading costs. Other aspects giving rise to trading costs, like depth or informational inefficiencies of prices, have not explicitly been included into our models. The reduced depth of stocks in case of narrower spreads shows that it is worth considering other aspects of the market structure influencing trading costs.

We saw in chapter 4 that market makers face costs from their activities that give rise to a spread to cover these costs. We showed the costs of market making to depend on the characteristics of the assets, its risk, the market participants, the share of informed investors, and the market makers, their risk aversion. These characteristics cannot directly be influenced by the structure of an exchange and hence it will be very difficult, if not impossible, to affect the costs of market making to reduce the spread.

Chapter 5 showed that under very general conditions market makers can coordinate their behavior such that even without reaching an explicit agreement, they quote spreads above their costs earning excess profits. Empirical evidence supports that such a behavior can be observed in actual markets. However, we also showed that in a realistic framework it will be very difficult to find direct evidence on implicit collusion. In the price setting behavior we found the properties of the spread not to differ sufficiently from the properties with competitive price setting and for the determination of the control structure the properties of the assets were found difficult to determine.

It is useful for exchanges to develop a market structure that breaks the implicit agreement between market makers and forces them to quote lower spreads that better reflect their costs of market making. In chapter 6 we pointed out some measures to reduce the effects arising from implicit collusion, although it will not be able to prevent such behavior completely. We proposed that increased competition from limit orders will directly reduce the spread faced by investors, although not necessarily those quoted by market makers. Other measures that should be considered are strict time priority for market makers to increase the incentives for breaking the agreement and anonymity of quotes set by market makers to limit competition on other factors than price. In several empirical investigations, experimental settings and theoretical models these measures have been shown to reduce the spread and therewith trading costs for investors.

It may be worth to consider in future research other dimensions of competition between market makers. Even if market makers do not perfectly compete on the price, they may compete in other dimensions, like the depth of the market.[1] The competition in these dimensions may also reduce trading costs. Empirical investigations, like Garvey and McCorry (1997), suggest that after the collusion between market makers on the NASDAQ had been detected, price competition increased, while competition in other dimensions reduced. The consequence according to their findings

[1] For the NYSE Doran (1999) gives evidence that members of the NYSE compete in several other dimensions than price, including order flow.

was that investors benefited only marginally from the reduced spread.

All considerations in this work only referred to dealer markets and the behavior of competing market makers. Other market forms, as briefly described in chapter 2.3, may be able to offer a market structure giving investors even lower costs than any dealer market could. Thus far there are only few contributions on the determination of the optimal market form, one we briefly considered in chapter 4.1.2. Here future research may bring further insights beyond the large number of empirical investigations comparing trading costs on stock exchanges with different market forms.[2] The results of these contributions, as well as a large number of empirical investigations, also suggest that in markets with limit orders dominating, like auction markets or the specialist system of the NYSE, trading costs are lower for investors. However, other aspects of the market structure, like anonymity or priority rules as well as asset characteristics, e.g. the share of informed investors, may influence these costs as well and bias empirical results. Therefore theoretical analysis may bring further insights into this aspect.

Recent developments changed the environment exchanges are operating in significantly. Nowadays exchanges face fierce competition from other trading facilities, exchanges as well as ECNs. Competition between these markets has only recently been considered in the academic literature, e.g. in Gehrig (1998), Foucault and Parlour (1999), Hendershott and Mendelson (2000), Huang (2000) or Benhamou and Serval (2000). Addressing this aspect in more detail can bring us further insights into other competitive forces limiting market power of colluding market makers and therewith reducing trading costs for investors, although the division of liquidity between competing markets may also increase them. Here also investigations into the advantages and disadvantages of a central order book, as proposed by the SEC for the United States, i.e. linking all markets electronically, seems to be very useful.[3]

[2] Recent contributions are Viswanathan and Wang (1999), Garfinkel and Nimalendran (1998), van Ness et al. (1999a), van Ness et al. (1999b) and He and Wu (1999).

[3] See Securities and Exchange Commission (2000b) and in more detail Securities and Exchange Commission (2000c).

Until now most exchanges are member-owned, i.e. the market makers and brokers that operate the market also own it. As has been pointed out in chapter 2.7, several exchanges are currently considering to become public companies. This change in the ownership structure may affect the behavior of brokers and market makers. Thus far only PIRRONG (1999) addresses the impact the ownership structure has on the behavior of market participants, future research can be expected to give us further clues on this issue.

Market participants take different roles over time. A market maker may also act as a broker, but he may also be an investor trading assets. It may be worth considering in future research how these different roles affect his behavior. Further extensions should also explicitly include the behavior of investors, e.g. their possibility to submit either market or limit orders as well as strategic behavior of informed investors.

We also have to keep in mind that we only considered the costs of investors in this work. It is obvious that any measures that are beneficial to investors by reducing the spread, reduce the profits made by market makers. Also the trading frequency and therewith the fees collected by brokers and the exchange will be affected as well as the need for stock exchanges to invest more into the extension of their trading facilities to handle increased trading volume. It will be useful in future research to address some of these aspects in more detail, e.g. in a general equilibrium model.

A final aspect may be worth considering; all models assume the market participants to operate under constant conditions. In contrast, recent developments show, the environment is changing rapidly. Therefore it should be worth to investigate the flexibility of market structures to adapt to these new conditions that cannot be foreseen. However, the literature thus far has not addressed this topic.

We have only been able to address some small aspects in market microstructure theory in this work, many aspects remain to be considered by future research.

Appendices

Appendix A

Utility theory

Since *John von Neumann* and *Oskar Morgenstern* introduced the expected utility hypothesis in 1944, it has become the most popular criterion for modeling decisions under risk. This appendix will give a brief introduction into the reasoning behind expected utility and its implications for risk aversion.

A.1 The expected utility hypothesis

The value, and therewith the returns of assets depend on their future cash flows. These future cash flows usually cannot be predicted with certainty by investors, they are a random variables, hence returns are also random variables. Investment decisions therewith have to be made under risk.[1] In their work VON NEUMANN AND MORGENSTERN (1953) presented a criterion to make an optimal decision if five axioms are fulfilled.[2]

[1] According to KNIGHT (1921, pp. 197ff.) a decision has to be made under *risk* if the outcome is not known with certainty, but the possible outcomes and the probabilities of each outcome are known. The probabilities can either be assigned by objective or subjective functions. KEYNES (1936, p. 68) defines risk as the possibility of the actual outcome to be different from the expected outcome. In contrast, under *uncertainty* the probabilities of each outcome are not known or even not all possible outcomes are known. CYMBALISTA (1998) provides an approach of asset valuation under uncertainty. In this work we only consider decisions to be made under risk.

[2] Many different ways to present these axioms can be found in the literature. We here follow the version of LEVY AND SARNAT (1972, p. 202)

Let $A = \{a_1, \ldots, a_N\}$ be the set of all possible alternatives an individual can choose between,[3] $S = \{s_1, \ldots, s_M\}$ all states that affect the outcome,[4] and $C = \{c_{11}, \ldots, c_{1M}; \ldots; c_{N1}, \ldots, c_{NM}\}$ the outcomes, where c_{ij} is the outcome if state s_j occurs and alternative a_i has been chosen.

Axiom 1. *A is completely ordered.*

A set is completely ordered if it is complete, i.e. we have either $a_i \succeq a_j$ or $a_j \succeq a_i$ for all $i, j = 1, \ldots, N$ and "$\succeq$" denotes the preference. The set has further to be transitive, i.e. if $a_i \succeq a_j$ and $a_j \succeq a_k$ we then have $a_i \succeq a_k$. Axiom 1 ensures that all alternatives can be compared with each other and are ordered consistently.[5]

To state the remaining axioms we have to introduce some notations. Let any alternative be denoted as a lottery, where each outcome c_{ij} has a probability of p_{ij}. We can write alternative i as

$$a_i = [p_{i1}c_{i1}, \ldots, p_{iM}c_{iM}] \text{ , where } \sum_{j=1}^{M} p_{ij} = 1 \text{ for all } i = 1, \ldots, N.$$

Axiom 2 (Decomposition of compound lotteries). *If the outcome of a lottery is itself a lottery (compound lottery), the first lottery can be decomposed into its final outcomes:*

Let $a_i = [p_{i1}b_{i1}, \ldots, p_{iM}b_{iM}]$ *and* $b_{ij} = [q_{ij1}c_1, \ldots, q_{ijL}c_L]$. *With* $p^*_{ik} = \sum_{l=1}^{M} p_{il}q_{ilk}$[6] *we have* $[p_{i1}b_{i1}, \ldots, p_{iM}b_{iM}] \sim [p^*_{i1}c_1, \ldots, p^*_{iM}c_L]$

Axiom 3 (Composition of compound lotteries). *If an individual is indifferent between two lotteries, they can be interchanged into a compound lottery:*

If $a_i = [p_{i1}b_{i1}, \ldots, p_{ij}b_{ij}, \ldots, p_{iM}b_{iM}]$ *and* $b_{ij} \sim [q_{ij1}c_1, \ldots, p_{ijL}c_L]$ *then*
$a_i \sim [p_{i1}b_{i1}, \ldots, p_{ij}[q_{ij1}c_1, \ldots, p_{ijL}c_L], \ldots, p_{iM}b_{iM}]$.

[3] As for $N = 1$ there is no decision to make for the individual it is required that $N \geq 2$.

[4] As with $M = 1$ the outcome can be predicted with certainty we need $M \geq 2$ possible states.

[5] The transitivity ensures consistent decisions of individuals. It is equivalent with the usual assumption in microeconomics that indifference curves do not cross. See SCHUMANN (1992, pp. 52 ff.).

[6] This representation of joint probabilities assumes that the two lotteries are independent. If the two lotteries where not independent the formula has to be changed, but the results remain valid. It is also assumed throughout this chapter that there is no joy of gambling, i.e. that there is no gain in utility from being exposed to uncertainty.

These two axioms ensure that lotteries can be decomposed into their most basic elements (axiom 2) and that more complex lotteries can be build up from their basic elements (axiom 3).

Axiom 4 (Monotonicity). *If two lotteries have the same two possible outcomes, then the lottery is preferred that has the higher probability on the more preferred outcome:*

Let $a_i = [p_{i1}c_1, p_{i2}c_2]$ *and* $b_i = [q_{i1}c_1, q_{i2}c_2]$ *with* $c_1 \succ c_2$, *if* $p_{i1} > q_{i1}$ *then* $a_i \succ b_i$.

given the same possible outcomes this axiom ensures the preference relation"$\succ$" to be a monotone transformation of the relation ">" between probabilities.

Axiom 5 (Continuity). *Let* a_i, b_i *and* c_i *be lotteries. If* $a_i \succ b_i$ *and* $b_i \succ c_i$ *then there exists a lottery* d_i *such that* $d_i = [p_1 a_i, p_2 c_i] \sim b_i$.

This axiom ensures the mapping from the preference relation "$\succ$" to the probability relation ">" to be continuous.

The validity of these axioms is widely accepted in the literature. Other axioms have been proposed, but the results from these axioms are identical to those to be derived in an instant.[7]

given these assumptions the following theorem can be derived, where U denotes the utility function.

Theorem 8 (Expected utility principle). *An alternative* a_i *will be preferred to an alternative* a_j *if and only if the expected utility of the former is larger, i.e*

$$a_i \succ a_j \Leftrightarrow E\left[U\left(a_i\right)\right] > E\left[U\left(a_j\right)\right].$$

Proof. Define a lottery $a_i = [p_{i1}c_1, \ldots, p_{iM}c_M]$, where without loss of generality $c_1 \succ c_2 \succ \cdots \succ c_M$. Such an order is ensured to exist by axiom 1.

[7] See MARKOWITZ (1959, p. xi).

Using axiom 5 we know that there exists a lottery such that

$$c_i \sim [u_{i1}c_1, u_{i2}c_M] = [u_i c_1, (1-u_i)c_M] \equiv c_i^*.$$

We can now use axiom 3 to substitute c_i by c_i^* in a_i:

$$a_i \sim [p_{i1}c_1^*, \ldots, p_{iM}c_M^*].$$

This alternative only has two possible outcomes: c_1 and c_M. By applying axiom 2 we get

$$a_i \sim [p_i c_1, (1-p_i)c_M]$$

with $p_i = \sum_{j=1}^{M} u_{ij}p_{ij}$, what is the definition of the expected value for discrete random variables: $p_i = E[u_i]$.[8]

The same manipulations as before can be made for another alternative a_j, resulting in

$$a_j \sim [p_j c_1, (1-p_j)c_M]$$

with $p_j = E[u_j]$. If $a_i \succ a_j$ then we find with axiom 4 that, as $c_1 \succ c_M$:

$$p_i > p_j.$$

The numbers u_{ij} we call the utility of alternative a_i if state s_j occurs. The interpretation as utility can be justified as follows: If $c_i \succ c_j$ then axiom 4 implies that $u_i > u_j$, we can use u_i to index the preference of the outcome, i.e. a higher u implies preference for this alternative and vice versa.

therewith we have shown that $a_i \succ a_j$ is equivalent to $E[U(a_i)] > E[U(a_j)]$. □

The criterion to choose between two alternatives, is to take that alternative with the highest expected utility. To apply this criterion the utility function has to be known. As in most cases we do not know the utility function, it is necessary to analyze this criterion further to derive a more handable criterion.

[8] The extension to continuous random variables is straightforward by replacing the probabilities with densities.

A.2 Risk aversion

> "Individuals are *risk averse* if they always prefer to receive a fixed payment to a random payment of equal expected value."[9]

From many empirical investigations it is known that individuals are risk averse, where the degree of risk aversion differs widely between individuals.[10] The Arrow-Pratt measure is the most widely used concept to measure this risk aversion. We will derive this measure following PRATT (1964), a similar measure has independently also been developed by ARROW (1963).

With the definition of risk aversion above, an individual prefers to receive a fixed payment of $E[x]$ to a random payment of x. To make the individual indifferent between a fixed payment and a random payment, there exists a number π, called *risk premium*, such that he is indifferent between receiving $E[x] - \pi$ and x. By applying the expected utility principle we see that the expected utility of these two payments has to be equal:

$$E\left[U\left(x\right)\right] = E\left[U\left(E\left[x\right] - \pi\right)\right] = U\left(E\left[x\right] - \pi\right). \tag{A.1}$$

The term $E\left[x\right] - \pi$ is also called the *cash equivalent* of x. Approximating the left side by a second order Taylor series expansion around $E\left[x\right]$ we get[11]

$$\begin{aligned} E\left[U\left(x\right)\right] &= E\Big[U\left(E\left[x\right]\right) + U'(E[x])(x - E[x]) \\ &\quad + \frac{1}{2}U''(E[x])(x - E[x])^2\Big] \\ &= U(E[x]) + U'(E[x])E[x - E[x]] \\ &\quad + \frac{1}{2}U''(E[x])E[(x - E[x])^2] \\ &= U(E[x]) + \frac{1}{2}U''(E[x])Var[x]. \end{aligned} \tag{A.2}$$

[9] DUMAS AND ALLAZ (1996, p. 30), emphasize added.

[10] Despite this clear evidence for risk aversion, many economic theories assume that individuals are risk neutral. Prominent examples in finance are the information-based models of market making (see section 5.5) and several asset pricing theories.

[11] We assume higher order terms to be negligible, what can be justified if x does not vary too much from $E\left[x\right]$.

where $U^{(n)}(E[x])$ denotes the nth derivative of U with respect to its argument evaluated at $E[x]$. In a similar way we can approximate the right side by a first order Taylor series around $E[x]$ and obtain

$$U(E[x] - \pi) = U(E[x]) + U'(E[x])\pi. \tag{A.3}$$

Inserting (A.2) and (A.3) into (A.1) and solving for the risk premium π we get

$$\pi = \frac{1}{2}\left(-\frac{U''(E[x])}{U'(E[x])}\right) Var(x). \tag{A.4}$$

PRATT (1964) now defines

$$z = -\frac{U''(E[x])}{U'(E[x])} \tag{A.5}$$

as the *absolute local risk aversion.* This can be justified by noting that the risk premia has to be larger the more risk averse an individual is and the higher the risk. The risk is measured by the variance of x, $Var[x]$,[12] hence the other term in (A.4) can be interpreted as risk aversion. Defining $\sigma^2 = Var[x]$ we get by inserting (A.5) into (A.4):

$$\pi = \frac{1}{2} z\sigma^2. \tag{A.6}$$

If we assume that individuals are risk averse we need $\pi > 0$, implying $z > 0$. It is reasonable to assume positive marginal utility, i.e. $U'(E[x]) > 0$, then this implies that $U''(E[x]) < 0$. This relation is also known as the first Gossen law and states the saturation effect.[13] The assumption of risk aversion is therefore in line with the standard assumptions in microeconomic theory.

The conditions $U'(E[x]) > 0$ and $U''(E[x]) < 0$ imply a concave utility function. The concavity of the function (radius) is determined by the risk aversion.[14]

Figure A.1 visualizes this finding for the simple case of two possible outcomes, x_1 and x_2, having equal probability of occurrence.

[12] A justification to use the variance as a measure of risk is given in appendix B.

[13] See SCHUMANN (1992, p. 49).

[14] For risk neutral individuals the risk premium, and hence the risk aversion, is zero, resulting in a zero second derivative of U, the utility function has to be linear. For risk loving individuals the risk premium and the risk aversion are negative, the second derivative of the utility function has to be positive, hence it is convex.

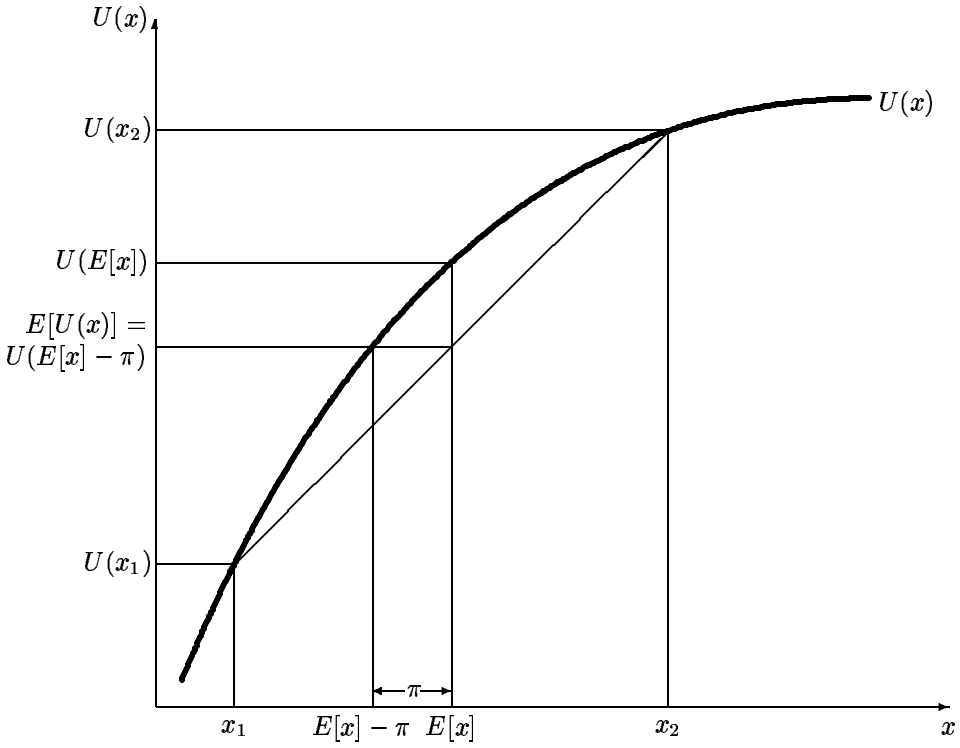


Figure A.1: The Arrow-Pratt measure of risk aversion

Appendix B

The portfolio selection theory

When considering to invest into asset markets, an investor has to make three decisions:

- the amount he wants to invest into the asset market,
- determine the assets he wants to invest in,
- determine the amount he wants to invest into each selected asset.

This appendix describes a method how to make these decisions and find an *optimal portfolio*.[1] Such a portfolio

> "... is more than a long list of good stocks and bonds. It is a balanced whole, providing the investor with protections and opportunities with respect to a wide range of contingencies. The investor should build toward an integrated portfolio which best suites his needs."[2]

For this reason the associated theory is called *portfolio selection theory* or short *portfolio theory*, rather than asset selection.[3] The portfolio selection theory has been developed by MARKOWITZ (1959), TOBIN (1958) and TOBIN (1966). Although the

[1] A *portfolio* is the entirety of all investments of an individual. See BÜSCHGEN (1991, p.552).
[2] MARKOWITZ (1959, p.3).
[3] See MARKOWITZ (1959, p.3).

concepts employed in their theory have much been criticized for capturing the reality only poorly, it has been the starting point for many asset pricing models and up to date there has been developed no widely accepted alternative.

B.1 The mean-variance criterion

Even by using the Arrow-Pratt measure of risk aversion, the utility function has to be known to determine the first and second derivative for basing a decision on the expected utility concept. Preferable would be a criterion that uses only observable variables instead of individual utility functions. For this purpose many criteria have been proposed,[4] the most widely used is the mean-variance criterion. Although it also is not able to determine the optimal decision, it restricts the alternatives to choose between by using the utility function.

The *mean-variance criterion* is the most popular criterion not only in finance. The reason is first that it is easy to apply and has some nice properties in terms of moments of a distribution and secondly by the use of this criterion in the basic works on portfolio selection by MARKOWITZ (1959), TOBIN (1958), and TOBIN (1966). Consequently, theories basing on their work, like the Capital Asset Pricing Model, also apply the mean-variance criterion, which by this mean became the most widely used criterion in finance.

It has the advantage that only two moments of the distribution of outcomes, mean and variance, have to be determined, whereas other criteria make use of the whole distribution.[5] The outcome is characterized by its expected value, the mean, and its risk, measured by the variance of outcomes.[6]

[4] See LEVY AND SARNAT (1972, ch. VII and ch. IX) for an overview of these criteria.

[5] See LEVY AND SARNAT (1972, pp. 307 ff.).

[6] One of the main critics of the mean-variance criterion starts with the assumption that risk can be measured by the variance. Many empirical investigations have shown that the variance is not an appropriate measure of risk. Many other risk measures have been proposed, see BRACHINGER AND WEBER (1997) for an overview, but these measures have the disadvantage of being less easily computable and difficult to implement as a criterion. In more recent models higher moments, such as skewness and kurtosis are also incorporated to cover the distribution in more detail.

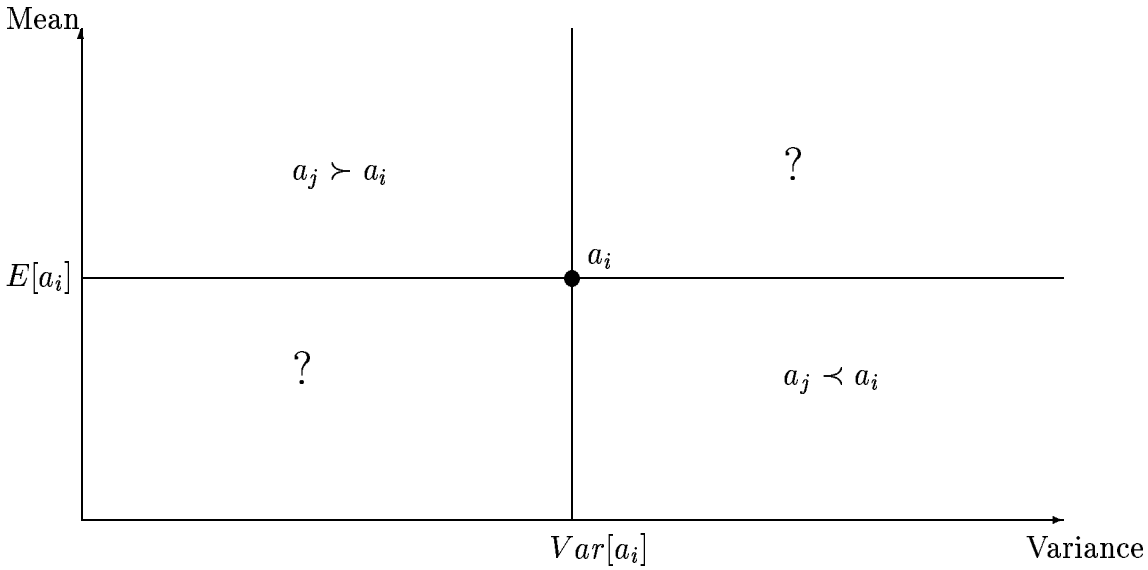


Figure B.1: The mean-variance criterion

The mean-variance criterion is defined as

$$a_i \succeq a_j \Leftrightarrow \begin{cases} Var[a_i] < Var[a_j] & \text{and} \quad E[a_i] \geq E[a_j] \\ \text{or} & \\ Var[a_i] \leq Var[a_j] & \text{and} \quad E[a_i] > E[a_j] \end{cases} . \tag{B.1}$$

It is a necessary, although not sufficient, condition to prefer a_i over a_j that $Var[a_i] \leq Var[a_j]$ and $E[a_i] \geq E[a_j]$. An alternative is preferred over another if it has a smaller risk (variance) and a larger mean. Nothing can in general be said about the preferences if $Var[a_i] > Var[a_j]$ and $E[a_i] > E[a_j]$, other decision rules have to be applied.[7]

In figure B.1 an alternative in the upper left and lower right areas can be compared to a_i by using the mean-variance criterion, while in the areas marked by "?" nothing can be said about the preferences. If we assume all alternatives to lie in a compact and convex set in the (μ, σ^2)-plane,[8] all alternatives that are not dominated by another alternative according to the mean-variance criterion lie on a line at the upper left of the set of alternatives. In figure B.2 this is illustrated where all alternatives are located in the oval. The undominated alternatives are represented by the bold

[7] See LEVY AND SARNAT (1972, pp. 308).

[8] We will see that this condition is fulfilled in the case of portfolio selection for all relevant portfolios.

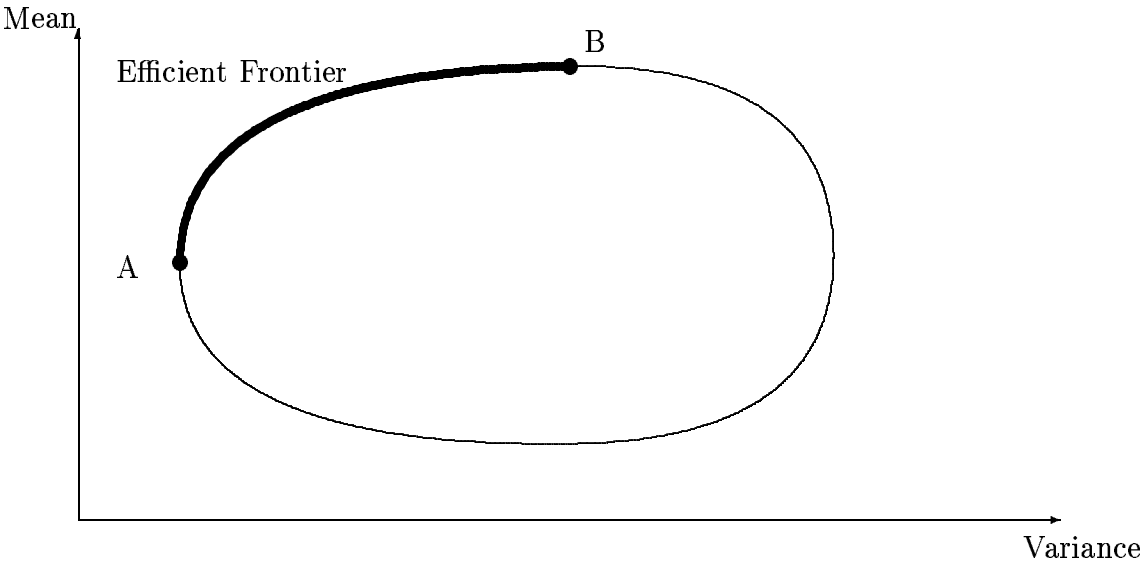


Figure B.2: The efficient frontier

line between points A and B. All alternatives that are not dominated by another alternative are called *efficient* and all efficient alternatives form the *efficient frontier.*[9] Without having additional information, e.g. the utility function, between efficient alternatives cannot be distinguished.

The mean-variance criterion can be shown to be not optimal in general, i.e. the true preferences are not always reflected by the results of this criterion.[10] If the utility function is quadratic, we will show that the mean-variance criterion always reflects the true preferences.[11]

Instead of defining the utility function by a term like $y = b_0 + b_1 x + b_2 x^2$ we can without loss of generality normalize the function by choosing $b_0 = 0$ and $b_1 = 1$.[12]

[9] See LEVY AND SARNAT (1972, pp. 318 ff.).

[10] See LEVY AND SARNAT (1972, pp. 310 f.). They also provide a generalization of the mean-variance criterion that is always optimal. As this criterion cannot be handled so easily, it is rarely applied and therefore not further considered here.

[11] See LEVY AND SARNAT (1972, pp. 379 ff.).

[12] The concept of expected utility implies that the utility function is only determined up to a positive linear transformation. This allows to apply the transformation $y \to \frac{y-b_0}{b_1}$ to achieve the normalization. See LEVY AND SARNAT (1972, pp. 205 and 379).

The utility function and its derivatives are therefore given by

$$U(x) = x + bx^2, \tag{B.2}$$

$$U'(x) = 1 + 2bx, \tag{B.3}$$

$$U''(x) = 2b. \tag{B.4}$$

According to (A.5) the Arrow-Pratt measure of risk aversion turns out to be

$$z = -\frac{2b}{1 + 2bE[x]}. \tag{B.5}$$

If we concentrate on risk averse individuals and assume reasonably positive marginal utility, (B.5) implies that

$$b < 0. \tag{B.6}$$

But if $b < 0$ we see from (B.3) that the marginal utility is only positive if

$$E[x] < -\frac{1}{2b}. \tag{B.7}$$

For large expected values the marginal utility can become negative. This unreasonable result can only be ruled out if the risk aversion is sufficiently small.[13]

If we define $E[x] = \mu$ and $Var[x] = \sigma^2$ we can write the expected utility as

$$E[U(x)] = E\left[x + bx^2\right] = \mu + bE\left[x^2\right] = \mu + b\left(\mu^2 + \sigma^2\right). \tag{B.8}$$

The indifference curves are obtained by totally differentiating both sides:

$$dE[U(x)] = (1 + 2b\mu)d\mu + 2\sigma d\sigma = 0. \tag{B.9}$$

The slope of the indifference curve in the (μ,σ)-plane is obtained by rearranging (B.9):[14]

$$\frac{d\mu}{d\sigma} = -\frac{2b\sigma}{1 + 2b\mu} = z\sigma > 0, \tag{B.10}$$

[13] This restriction on the expected value is another argument often brought forward against the use of a quadratic utility function and hence the mean-variance criterion. Another argument is that the risk aversion increases with the expected outcome: $\frac{\partial z}{\partial E[x]} = \frac{4b^2}{(1+2bE[x])^2} > 0$, which contradicts empirical findings. Moreover in many theoretical models a constant risk aversion is assumed, which has been shown by Pratt (1964) to imply an exponential utility function. If the expected outcome does not vary too much, constant risk aversion can be approximated by using a quadratic utility function.

[14] Instead of using the variance as a measure of risk, it is more common to use its square root, the standard deviation. As the square root is a monotone transformation, the results are not changed by this manipulation.

i.e. for risk averse investors the indifference curves have a positive slope in the (μ,σ)-plane.

The equation of the indifference curve is obtained by solving (B.8) for μ:[15]

$$\begin{aligned} E[U(x)] &= \mu + b\mu^2 + b\sigma^2 \\ \mu^2 + \frac{1}{b}\mu + \sigma^2 &= \frac{E[U(x)]}{b} \\ \left(\mu + \frac{1}{2b}\right)^2 + \sigma^2 &= \frac{1}{b}E[U(x)] + \frac{1}{4b^2}. \end{aligned} \tag{B.11}$$

Defining $r^* = -\frac{1}{2b}$ as the expected outcome that must not be exceeded for the marginal utility to be positive according to equation (B.7), we can rewrite the equation for the indifference curves as

$$(\mu - r^*) + \sigma^2 = -2r^*E[U(x)] + r^{*2} \equiv R^2, \tag{B.12}$$

which is the equation of a circle with center $\mu = r^*$, $\sigma = 0$ and radius R.[16] With this indifference curve, which has as the only parameter a term linked to the risk aversion, it is now possible to determine the optimal alternative out of the efficient alternatives, that is located at the point where the efficient frontier is tangential to the indifference curve. Figure B.3 shows the determination of the optimal alternative C.

We will now show that with a quadratic utility function the mean-variance criterion is optimal.[17] We assume two alternatives with $\mu_i = E[a_i] > E[a_j] = \mu_j$. Let further $\sigma_i^2 = Var[a_i]$ and $\sigma_j^2 = Var[a_j]$. If $a_i \succ a_j$ it has to be shown that

$$E[U(a_i)] > E[U(a_j)].$$

Substituting the utility functions gives

$$\mu_i + b\mu_i^2 + b\sigma_i^2 > \mu_j + b\mu_j^2 + b\sigma_j^2,$$

[15] See SHARPE (1970, pp. 198 f.)

[16] The results that the indifference curves are circles gives rise to another objection against the use of a quadratic utility function. An individual with a quadratic utility function should be indifferent between an expected outcome of $r^* + v$ and $r^* - v$ for any given σ. From the mean-variance criterion (B.1) we know that for a given σ the alternative with the higher expected outcome will be preferred. In practice this problem is overcome by using only the lower right sector of the circle.

[17] Such a proof is given e.g. in LEVY AND SARNAT (1972, pp. 385 ff.).

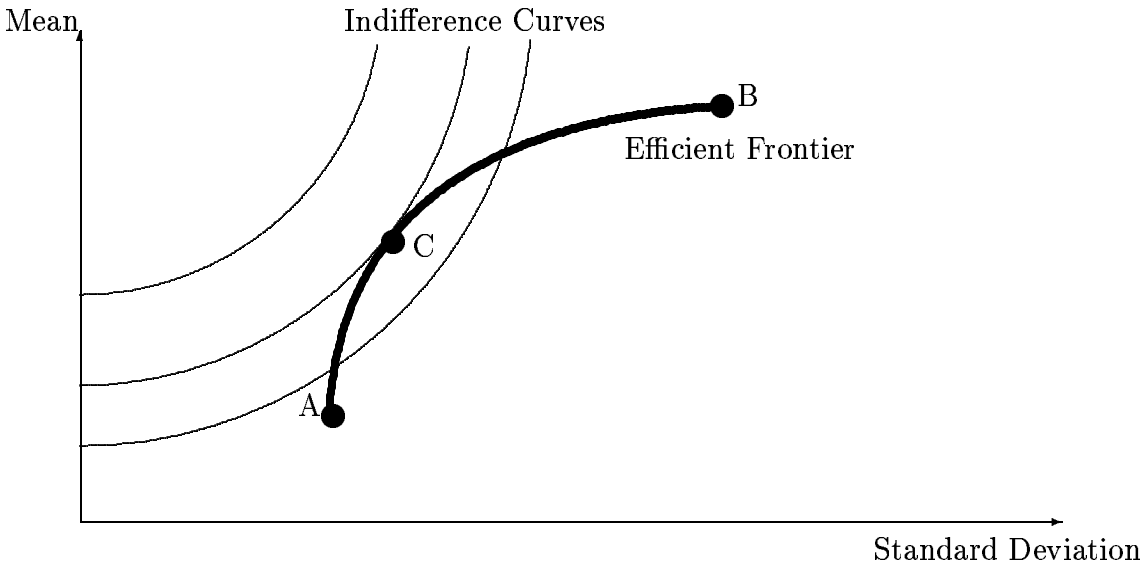


Figure B.3: Determination of the optimal alternative

$$\mu_i - \mu_j + b(\mu_i^2 - \mu_j^2) + b(\sigma_i^2 - \sigma_j^2) = (\mu_i - \mu_j)\left[1 + b(\mu_i + \mu_j)\right] + b(\sigma_i^2 - \sigma_j^2) > 0.$$

Dividing by $-2b > 0$ gives us

$$(\mu_i - \mu_j)\left[-\frac{1}{2b} - \frac{\mu_i + \mu_j}{2}\right] - \frac{\sigma_i^2 - \sigma_j^2}{2} > 0. \tag{B.13}$$

From (B.7) we know that $-\frac{1}{2b} > \mu_i$ and $-\frac{1}{2b} > \mu_j$, hence we find that

$$-\frac{1}{2b} > \frac{\mu_i + \mu_j}{2} \tag{B.14}$$

With the assumption that $\mu_i > \mu_j$ and as $b < 0$ the first term in (B.13) is positive. If now $\sigma_i^2 \leq \sigma_j^2$ as proposed by the mean-variance criterion equation (B.13) is fulfilled and we have shown that it represents the true preferences.

If $\sigma_i^2 > \sigma_j^2$ in general nothing can be said which alternative will be preferred. For $\mu_i = \mu_j$ we need $\sigma_i^2 < \sigma_j^2$ in order to prefer a_i over a_j. This is exact the statement made by the mean-variance criterion in (B.7). Therewith it has been shown that in the case of a quadratic utility function the mean-variance criterion is optimal, i.e. represents the true preferences.[18]

[18] A quadratic utility function is not only a sufficient condition for the optimality of the mean-variance criterion, but also a necessary condition. This is known in the literature as the Schneeweiss-Theorem.

B.2 The Markowitz frontier

The portfolio selection theory is based on several assumptions:[19]

- no transaction costs and taxes,
- assets are indefinitely divisible,
- each investor can invest into every asset without restrictions,
- investors maximize expected utility by using the mean-variance criterion,
- prices are given and cannot be influenced by investors (competitive prices),
- the model is static, i.e. only a single time period is considered.

Some of these assumptions, like the absence of transaction costs and taxes have been lifted by more recent contributions without giving fundamental new insights.

In portfolio selection theory the different alternatives to choose between are the compositions of the portfolios, i.e. the weight each asset has.[20] Assume an investor has to choose between $N > 1$ assets, assigning a weight of x_i to each asset. The expected return of each asset is denoted μ_i and the variance of the returns by $\sigma_i^2 > 0$ for all $i = 1, \ldots, N$.[21] The covariances between two assets i and j will be denoted σ_{ij}.

The weights of the assets an investor holds, have to sum up to one and are assumed to be positive as we do not allow for short sales at this stage:

$$\sum_{i=1}^{N} x_i = 1, \tag{B.15}$$
$$x_i \geq 0, \; i = 1, \ldots, N.$$

[19] See LINTNER (1965, p. 15).

[20] The decision which portfolio is optimal does not depend on total wealth for a given constant risk aversion, hence it can be analyzed by dealing with weights only. See LEVY AND SARNAT (1972, pp. 420 f.).

[21] Instead of investigating final expected wealth and its variance after a given period of time (the time horizon), we can use the expected return and variances of returns as they are a positive linear transformation of the wealth. As has been noted above, the decision is not influenced by such a transformation when using expected utility.

For the moment assume that there are only $N = 2$ assets. The characteristics of each asset can be represented as a point in the (μ,σ)-plane. We then can derive the location of any portfolio in the (μ,σ)-plane by combining these two assets.[22]

The expected return and the variance of the return of the portfolio is given by

$$\text{(B.16)} \qquad \mu_p = x_1\mu_1 + x_2\mu_2 = \mu_2 + x_1(\mu_1 - \mu_2),$$

$$\text{(B.17)} \qquad \begin{aligned} \sigma_p^2 &= x_1^2\sigma_1^2 + x_2^2\sigma_2^2 + 2x_1x_2\sigma_{12} \\ &= \sigma_2^2 + x_1^2(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}) + 2x_1(\sigma_1\sigma_2\rho_{12} - \sigma_2^2), \end{aligned}$$

where $\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ denotes the correlation of the two assets.

The portfolio with the lowest risk is obtained by minimizing (B.17). The first order condition is

$$\frac{\partial\sigma_p^2}{\partial x_1} = 2x_1(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}) + 2(\sigma_1\sigma_2\rho_{12} - \sigma_2^2) = 0.$$

The second order condition for a minimum is fulfilled unless $\sigma_1 = \sigma_2$ and $\rho_{12} \neq 1$:

$$\frac{\partial^2\sigma_p^2}{\partial x_1^2} = 2(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}) > 2(\sigma_1 - \sigma_2)^2 > 0$$

Solving the first order condition gives the weights in the *minimum risk portfolio* (*MRP*):

$$\text{(B.18)} \qquad x_1^{MRP} = \frac{\sigma_2^2 - \sigma_1\sigma_2\rho_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}}.$$

The minimum variance can be obtained by inserting (B.18) into (B.17):

$$\text{(B.19)} \qquad \begin{aligned} \sigma_{MRP}^2 &= \sigma_2^2 + \frac{(\sigma_2^2 - \sigma_1\sigma_2\rho_{12})^2}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}} - 2\frac{(\sigma_2^2 - \sigma_1\sigma_2\rho_{12})^2}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}} \\ &= \sigma_2^2 - \frac{(\sigma_2^2 - \sigma_1\sigma_2\rho_{12})^2}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}} \\ &= \frac{\sigma_1^2\sigma_2^2(1 - \rho_{12}^2)}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}}. \end{aligned}$$

If the returns of the two assets are uncorrelated ($\rho_{12} = 0$), then (B.19) reduces to

$$\text{(B.20)} \qquad \sigma_{MRP}^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

[22] See TOBIN (1966, pp. 22ff.).

This variance is smaller than the variance of any of these two assets.[23] By holding an appropriate portfolio, the variance, and hence the risk, can be reduced, whereas the expected return lies between the expected returns of the two assets.

With perfectly negative correlated assets ($\rho_{12} = -1$) we find that

$$\sigma^2_{MRP} = 0 \tag{B.21}$$

and the risk can be eliminated from the portfolio.

In the case of perfectly correlated assets ($\rho_{12} = 1$) the minimum variance is the variance of the asset with the lower variance:

$$\sigma^2_{MRP} = \begin{cases} \sigma_1^2 & \text{if} \quad \sigma_1^2 \leq \sigma_2^2 \\ \sigma_2^2 & \text{if} \quad \sigma_1^2 > \sigma_2^2 \end{cases} . \tag{B.22}$$

We can derive a general expression for the mean-variance relation:

$$\begin{aligned} \sigma_p^2 - \sigma^2_{MRP} &= \sigma_2^2 + x_1^2(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}) + 2x_1(\sigma_1\sigma_2\rho_{12} - \sigma_2^2) \\ &\quad - \frac{\sigma_1^2\sigma_2^2(1-\rho_{12}^2)}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}} \\ &= (x_1 - x_1^{MRP})^2(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}). \end{aligned} \tag{B.23}$$

With μ_{MRP} denoting the expected return of the minimum risk portfolio, we find that

$$\mu_p - \mu_{MRP} = (x_1 - x_1^{MRP})(\mu_1 - \mu_2). \tag{B.24}$$

Solving (B.24) for $x_1 - x_1^{MRP}$ and inserting into (B.23) we obtain after rearranging:

$$(\mu_p - \mu_{MPR})^2 = \frac{\sigma_p^2 - \sigma^2_{MRP}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}}(\mu_1 - \mu_2)^2. \tag{B.25}$$

This equation represents a hyperbola with axes[24]

$$\begin{aligned} \mu_p &= \mu_{MRP} + \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}}}\sigma_p, \\ \mu_p &= \mu_{MRP} - \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}}}\sigma_p. \end{aligned}$$

[23] Suppose $\sigma^2_{MRP} = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2} > \sigma_2^2$, this would imply that $\sigma_1^2 > \sigma_1^2 + \sigma_2^2$ and hence $\sigma_2^2 < 0$, which contradicts the assumption that $\sigma_2^2 > 0$. A similar argument can be used to show that $\sigma^2_{MRP} < \sigma_1^2$.

[24] See TOBIN (1966, p.30).

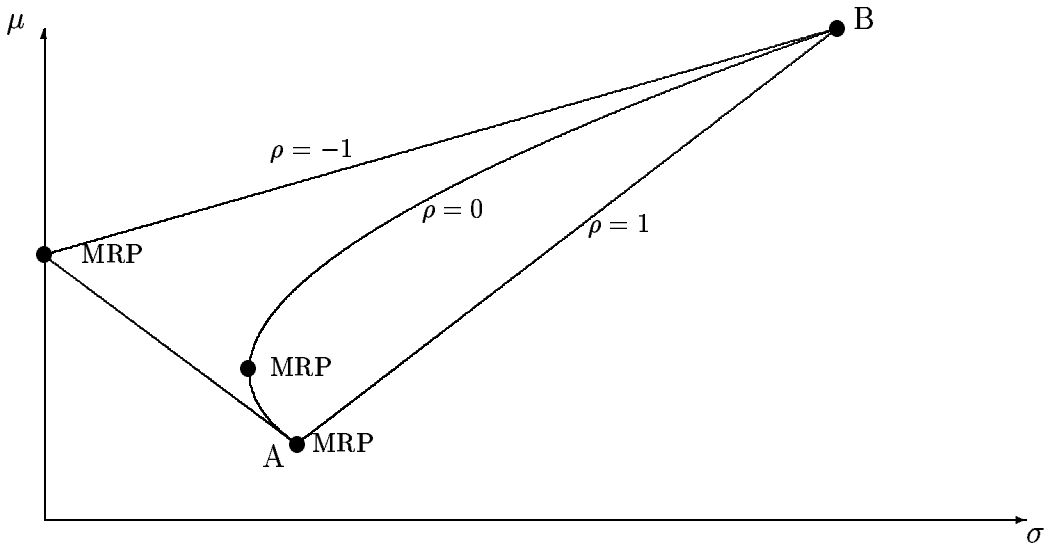


Figure B.4: Efficient portfolios with two assets

The efficient portfolios lie on the upper branch of this hyperbola, i.e. above the minimum risk portfolio.[25] Figure B.4 shows the efficient portfolios for different correlations. It can easily be shown that in the case of perfect positive correlation the efficient portfolios are located on a straight line connecting the two assets, in case of perfectly negative correlation on straight lines connecting the assets with the minimum risk portfolio. Between efficient portfolios can only be distinguished by using the utility function. Figure B.5 adds the indifference curve to the opportunity locus and determines the location of the *optimal portfolio* (OP). The location of the optimal portfolio depends on the risk aversion of the investor, the more risk averse the investor is the more close the optimal portfolio will be located to the minimum risk portfolio.

It is now possible to introduce a third asset. In a similar way hyperbolas can be deducted representing all combinations of this asset with one of the other two. Furthermore we can view any portfolio consisting of the two other assets as a single new asset and can combine it in the same manner with the third asset. Figure B.6 illustrates this situation. All achievable portfolio combinations are now located in

[25] The efficient frontier is also called *opportunity locus*.

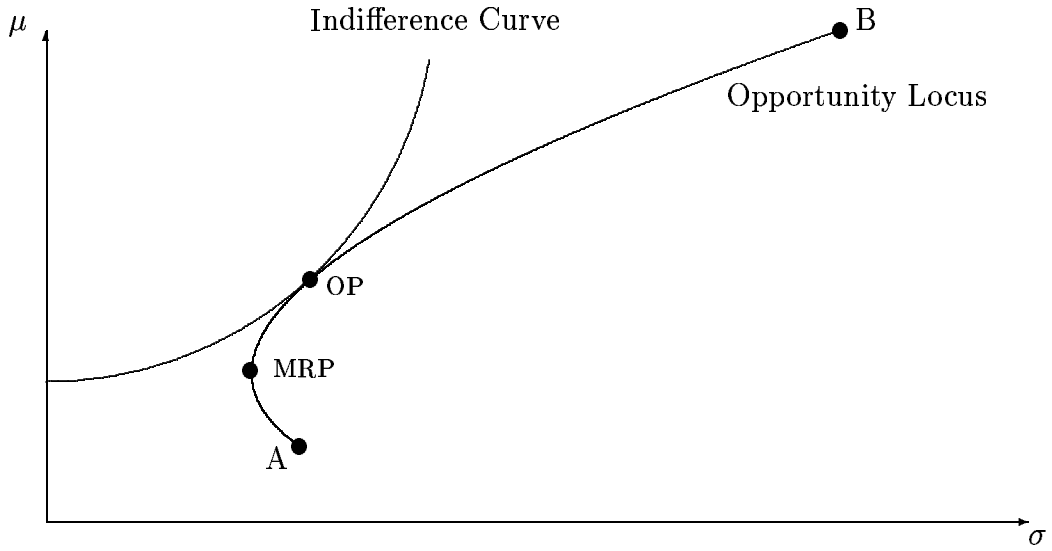


Figure B.5: Determination of the optimal portfolio with two assets

the area bordered by the bold line connecting points A and C, where the bold line encircling the different hyperbolas is the new opportunity locus.

This concept can be generalized to $N > 3$ assets in the same manner. All achievable assets will be located in an area and the efficient frontier will be a hyperbola. Using the utility function the optimal portfolio can be determined in a similar way as in the case of two assets as shown in figure B.7. If an asset is added, the area of achievable portfolios is enlarged and encompasses the initial area. This can simply be proved by stating that the new achievable portfolios encompass also the portfolios assigning a weight of zero to the new asset. With a weight of zero these portfolios are identical to the initially achievable portfolios. To these portfolios those have to be added assigning a non-zero weight to the new asset. Therefore the efficient frontier moves further outward to the upper left. By adding new assets the utility can be increased.

Thus far it has been assumed that $\sigma_i^2 > 0$, i.e. all assets were risky. It is also possible to introduce a riskless asset, e.g. a government bond, with a variance of zero. Define the return of the riskless asset by r, then in the case of two assets we

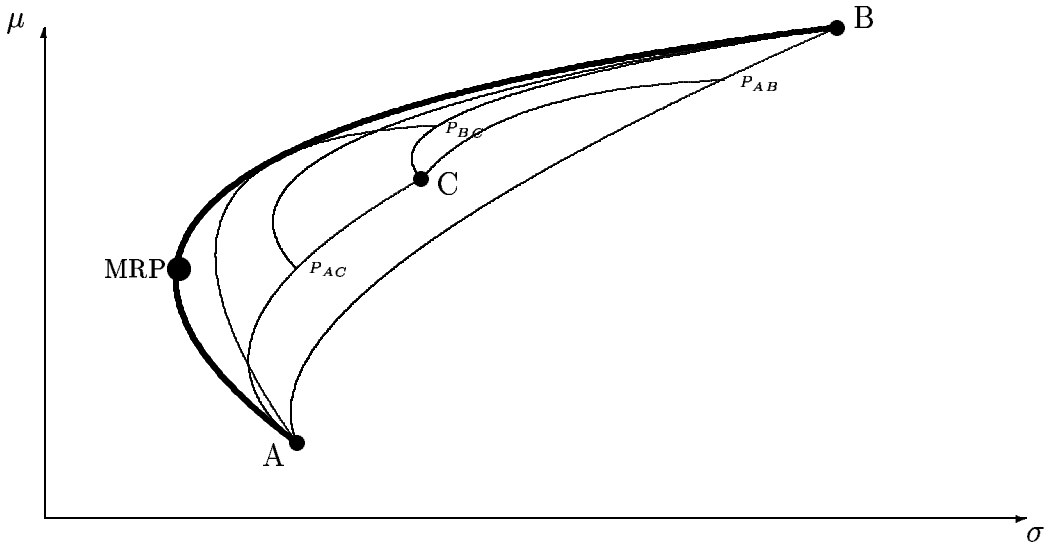


Figure B.6: Portfolio selection with three assets

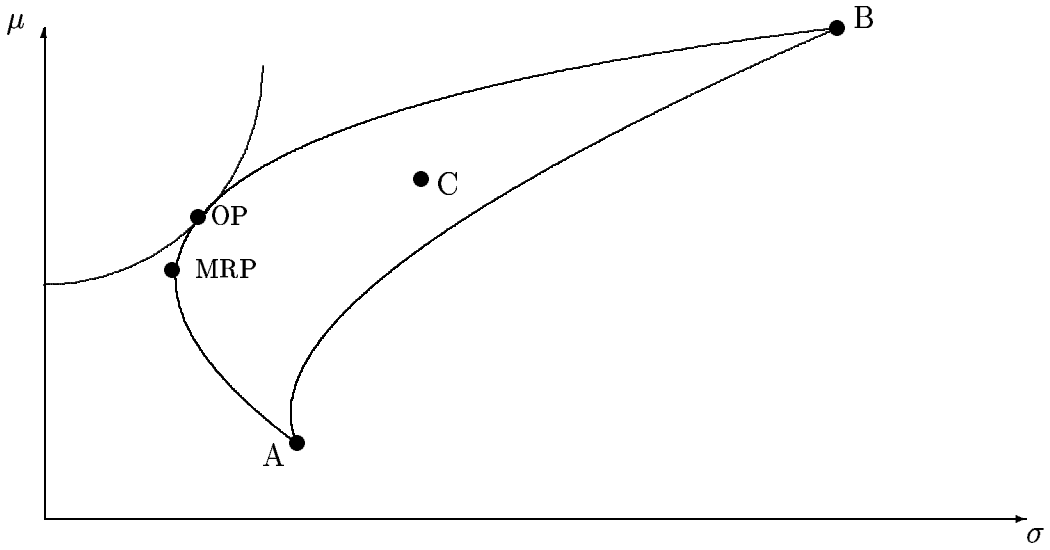


Figure B.7: The optimal portfolio with $N > 2$ assets

get from (B.16) and (B.17):[26]

$$\mu_p = x_1\mu_1 + x_2 r = r + x_1(\mu_1 - r), \tag{B.26}$$

$$\sigma_p^2 = x_1^2\sigma_1^2. \tag{B.27}$$

Solving (B.27) for x_1 and inserting into (B.26) gives

$$\mu_p = r + \frac{\sigma_p}{\sigma_1}(\mu_1 - r) = r + \frac{\mu_1 - r}{\sigma_1}\sigma_p. \tag{B.28}$$

The expected return of the portfolio is linear in the variance of the portfolio return, i.e. the hyperbola reduces to a straight line from the location of the riskless asset, $(0, r)$, to the location of the asset. In the case of many risky assets we can combine every portfolio of risky assets with the riskless asset and obtain all achievable portfolios. As shown in figure B.8 all achievable portfolios are located between the two straight lines, the upper representing the efficient frontier. There exists a portfolio consisting only of risky assets that is located on the efficient frontier. It is the portfolio consisting only of the risky assets at which the efficient frontiers with and without a riskless asset are tangential.[27] This portfolio is called the *optimal risky portfolio* (*ORP*). The efficient frontier with a riskless asset is also called the *capital market line.*

All efficient portfolios are located on the capital market line, consequently they are a combination of the riskless asset and the optimal risky portfolio. The optimal portfolio can be obtained in the usual way by introducing the indifference curves. As the optimal portfolio always is located on the capital market line, it consists of the risky asset and the optimal risky portfolio. Which weight is assigned to each depends on the risk aversion of the investor, the more risk averse he is the more weight he will put on the riskless asset. The weights of the optimal risky portfolio do not depend on the risk aversion of the investor. The decision process can therefore be separated into two steps, the determination of the optimal risky portfolio and then the determination of the optimal portfolio as a combination of the ORP with

[26] See TOBIN (1958).

[27] It is also possible that no tangential point exists, in this case a boundary solution exists and the risky portfolio consists only of a single risky asset.

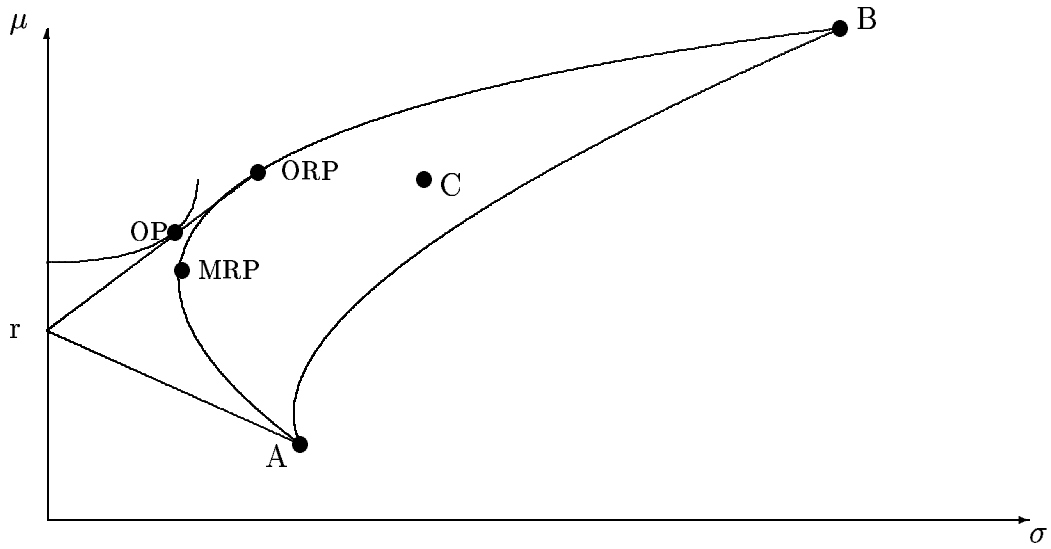


Figure B.8: The optimal portfolio with a riskless asset

the riskless asset. As this result has first been presented by TOBIN (1958) it is also called the *Tobin separation theorem.*[28]

So far we have assumed that $x_i \geq 0$ for all $i = 1, \ldots, N$. If we allow now some x_i to be negative the possibilities to form portfolios is extended. An asset with $x_i < 0$ means that the asset is sold short, i.e. it is sold without having owned it before. This situation can be viewed as a credit that has not been given and has not to be repaid in money (unless money is the asset), but in the asset. The assets can be the riskless asset or the risky assets, in the former case the short sale is an ordinary credit. It is assumed that credits can be obtained at the same conditions (interest rate or expected return and risk) as investing in the asset.

By allowing short sales the efficient frontier of the risky portfolios further moves to the upper left as new possibilities to form portfolios are added by lifting the restriction that the weights must be non-negative. Therewith the capital market line

[28] For investors being less risk averse it is possible that the optimal portfolio is located on the part of the efficient frontier above the ORP, in this case the optimal portfolio does not contain the riskless asset and assigns different weights to the risky assets compared to the ORP. Therefore in general the Tobin separation theorem does only hold with the inclusion of short sales, as described in the next paragraph.

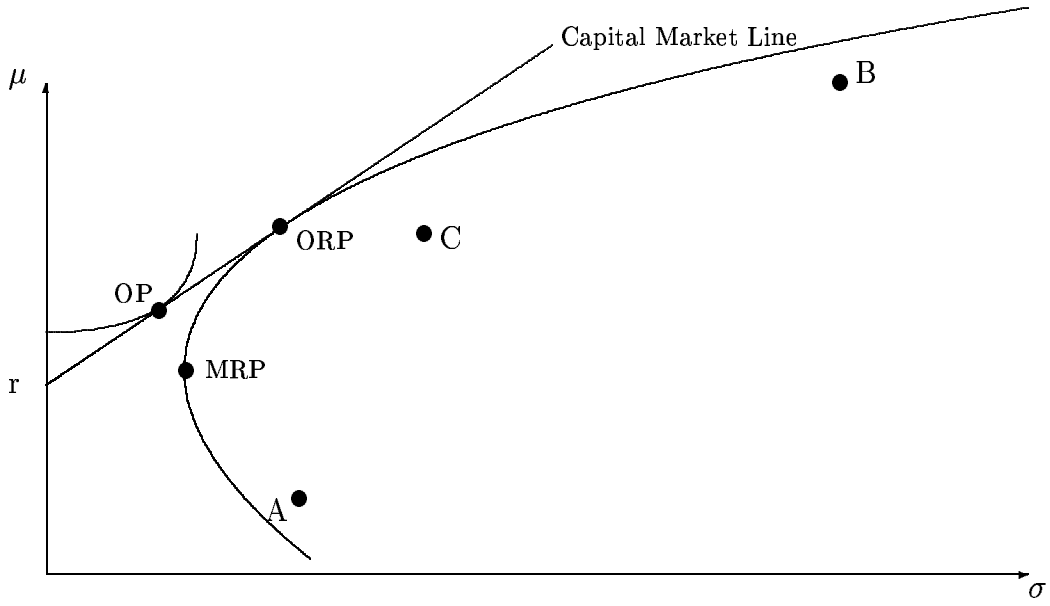


Figure B.9: Portfolio selection with short sales

becomes steeper and the utility of the optimal portfolio increases. Figure B.9 illustrates this case. The Capital Market Line extends beyond the ORP and therewith the optimal portfolio will always be a combination of the riskless asset and the ORP. The Tobin separation theorem applies in all cases, independent of the degree of risk aversion. If the ORP is located above the ORP, the riskless asset is sold short and a larger fraction of the optimal portfolio consists of the ORP.

In applying the portfolio theory to determine the optimal portfolio several problems are faced:

- determination of the risk aversion of the investor,
- determination of the expected returns, variances and covariances of the assets,
- computation of the efficient frontier and the optimal portfolio.

There exists no objective way to determine the risk aversion of an investor, most investors are only able to give a qualitative measure of their risk aversion, if at

all. The transformation into a quantitative measure is an unsolved, but for the determination of the optimal portfolio critical problem. It is important for the allocation between the riskless asset and the optimal risky portfolio.

Expected returns, variances and covariances can be obtained from estimates based on past data. But there is no guarantee that these results are reasonable for the future. It is also possible to use other methods to determine these moments, e.g. by using subjective beliefs. The determination of these moments are critical for the determination of the optimal risky portfolio.

To determine the efficient frontier and the optimal portfolio non-trivial numerical optimization routines have to be applied.[29] Advances in computer facilities and the availability of these routines do not impose a threat anymore as it has done in former years.

When having solved the above mentioned problems, the portfolio theory does allow to answer the questions raised at the beginning of this appendix:

- the share to be invested into risky assets is determined by the optimal portfolio,
- the assets to invest in are those included in the optimal risky portfolio,
- the shares to invest in each selected asset are given by the weights of the optimal risky portfolio.

The portfolio theory has developed a method how to allocate resources optimal. Although mostly only financial assets are included, other assets like human capital, real estate and others can easily be included, although it is even more difficult to determine their characteristics.

A shortcoming of the portfolio theory is that it is a static model. It determines the optimal portfolio at a given date. If the time horizon is longer than one period,

[29] For a detailed description of the mathematical concepts to solve these problems see MARKOWITZ (1959) and ASCHINGER (1990).

the prices of assets change over time, and therewith the weights of the assets in the initial portfolio change. Even if the expected returns, variances and covariances do not change, this requires to rebalance the portfolio every period. As assets with a high realized return enlarge their weight, they have partially to be sold to buy assets which had a low return (sell the winners, buy the losers). In a dynamic model other strategies have been shown to achieve a higher expected utility for investors, but due to the static nature of the model such strategies cannot be included in this framework.

Appendix C

The Rational Expectations Approach

This appendix introduces the concept of rational expectations. The literature on this subject has grown rapidly since it was first introduced by MUTH (1961). Therefore here only the basic concept is presented, it is not the aim to give an overview of the entire theory. The first applications of rational expectations could be found in macroeconomics in the 1970's, most prominently by the Nobel laureate Robert E. Lucas. Nowadays rational expectations are used in nearly every field of economics, especially in financial markets theory.[1]

Using his individual information, every investor will form his expectations on a certain economic variable which is not known with certainty, e.g. the fundamental value of an asset and act on this individual expectation. These expectations and actions transform into prices, where the price realized may be different from the expectations of all investors. In order to analyze how expectations and the therewith associated actions transform into prices it is necessary to model the way expectations are formed.

C.1 The rational expectations hypothesis

When introducing *rational expectations* MUTH (1961, p. 316) proposed that expectations

[1] SHEFFRIN (1996) presents several examples of using rational expectations in different fields of economics.

"... are essentially the same as the predictions of the relevant economic theory."

He explains this definition of rational expectations by stating that these expectations tend to be distributed for a given information set about the distribution of the theory,[2] i.e. on average the mean of the expectations equals the forecast of the theory.[3] This implicitly states that all investors agree on the theory, i.e. on the structure of the economy as well as on the parameters. The assumption that rational expectations produce no systematic errors, implies that investors do not only have to agree on the theory, but that the theory also has to be true.

Rational expectations are represented by conditional expectations. Let X_t be the variable to be forecasted and Ω_{t-1} the information available, then with X_t^e representing the expectation of X_t we define rational expectations as

$$X_t^e = E\left[X_t|\Omega_{t-1}\right]. \tag{C.1}$$

That there exists no systematic error in these expectations gives

$$E\left[X_t - X_t^e|\Omega_{t-1}\right] = 0, \tag{C.2}$$

$$E\left[(X_t - X_t^e)X_t|\Omega_{t-1}\right] = 0, \tag{C.3}$$

i.e. the expected forecast error is zero and uncorrelated with the true value of the variable. Many examples show that this in general only holds if the theory is correct.[4]

A *narrow version* of the definition given by MUTH (1961) has been presented by LUCAS AND PRESCOTT (1971). They require not only the mean of the expectations and the predictions of the theory to coincide, but they must have the same probability distribution.

[2] See MUTH (1961, p. 316).
[3] See REDMAN (1992, p.7).
[4] See e.g. PESARAN (1987, pp. 29 ff.).

A *weak version* assumes that information is costly and therefore every individual only acquires information until his marginal costs and benefits equal. Only this information acquired is used to form rational expectations, not all available information. It also allows for rational expectations that are not based on the correct theory. Investors then form their expectations "as if" they knew the correct model. These expectations in general will be biased, i.e. not fulfill equations (C.2) and (C.3).[5]

Rational expectations have much been criticized for the very restrictive assumptions, especially in the narrow version. The assumption that all individuals know the true model of the economy has been the main target, but also the assumption that all available information has to be taken into account.[6] The weak version has also been attacked as by its definition it allows to define every form of expectations to be called rational.[7] Especially how investors could learn the true model of the economy remains an unsolved problem.[8]

A more general criticism on rational choice theory, and therewith rational expectations, in most cases addresses the fact that individuals are not only rational but in many situations behave emotional or imitate others. These behavioral approaches have attracted increased attention in recent years, especially in finance, where they are used to explain several of features of asset prices that cannot be explained using rational behavior.[9] In most cases models using rational choice make use of very restrictive assumptions to be able to derive results by using sophisticated mathematical methods.[10] These assumptions make predictions of behavior in actual markets very difficult and in many cases they fail. Here BEED AND BEED (2000) question

[5] See PESARAN (1987, p.23). He also points out that in such situations self-fulfilling prophecies can occur until the model is identified to be wrong and a sudden change in the model applied may correct the situation.

[6] See REDMAN (1992, pp. 13 ff.).

[7] See GOMES (1982, p. 52).

[8] See also FRIEDMAN (1979) on this point.

[9] See THALER (1993) or SHLEIFER (2000) for an overview.

[10] Another objective against the current economic theory and its dominating rational choice theory is the concentration on self defined and very abstract problems rather than on "real world problems" that concern the society as FREY (2000, 25 f.) points out.

the usefulness of rational choice theory and its contribution to the advancement of economic knowledge in general. However, at present there has been developed no more powerful tool to address economic problems.

By allowing investors to learn some aspects of the economy, the narrow version can partly be weakened. We can allow investors to learn some parameters of the economy, e.g. the beta in the CAPM. If we assume that all investors know the structure of the economy, but not the parameters, it can be shown that they will learn the true parameters over time, i.e. their expectations converge to the narrow version. The most widely used concept of learning is Bayesian learning to be presented in the next section.

Despite these modifications it remains an unsolved problem how investors learn the true structure of the economy. As no other theory on expectation formation exists that provides a better explanation of economic phenomena, rational expectations are widely used in economics.

C.2 Bayesian learning

We assume investors to know the structure of the economy, but not the parameters in the model. In a first step they form beliefs about these parameters, e.g. by assigning a random number or another reasoning. Every investor can have different beliefs. Based on these first beliefs they form their expectations and act accordingly. When they see the realization of the outcome, they realize that their expectations were not correct. As they know the structure of the economy, the only source of these deviations can be the values of the parameters assigned. Hence they want to change their beliefs about these parameters. This process continues until expectations and realizations coincide and the beliefs are correct.[11]

[11] See PESARAN (1987, pp. 32 f.).

If there are random variables in the economy, the most widely used method to model changes in beliefs uses Bayes rule and hence is called *Bayesian learning.*[12]

We know from probability theory that for discrete random variables with Prob denoting the probability:[13]

$$\text{(C.4)} \quad \text{Prob}[\Omega_t] = \text{Prob}[\Omega_t | a = a_t]\text{Prob}[a = a_t] + \text{Prob}[\Omega_t | a \neq a_t]\text{Prob}[a \neq a_t],$$

$$\text{(C.5)} \quad \text{Prob}[a = a_t | \Omega_t] = \frac{\text{Prob}[a = a_t, \Omega_t]}{\text{Prob}[\Omega_t]},$$

$$\text{(C.6)} \quad \text{Prob}[\Omega_t | a = a_t] = \frac{\text{Prob}[a = a_t, \Omega_t]}{\text{Prob}[a = a_t]},$$

where a is the true parameter and a_t the current belief of the value of this parameter. Combining relations (C.4) - (C.6) gives

(C.7)

$$\text{Prob}[a = a_t | \Omega_t] = \frac{\text{Prob}[\Omega_t | a = a_t]\text{Prob}[a = a_t]}{\text{Prob}[\Omega_t | a = a_t]\text{Prob}[a = a_t] + \text{Prob}[\Omega_t | a \neq a_t]\text{Prob}[a \neq a_t]},$$

which is also known as *Bayes rule.* The probability that $a = a_t$ in the last period, $\text{Prob}[a = a_t]$, is called the *prior belief.* On this belief the investor based his decision. given this belief for all possible values the entire distribution is known. In the next period he learns the realization of the process, that forms part of his new information set, Ω_t. Based on this new information he changes his belief that $a = a_t$ to $\text{Prob}[a = a_t | \Omega_t]$ according to equation (C.7), his *posterior belief.* For the next period these beliefs are his prior beliefs, which he updates on the new information received in the next period. Applying this relation for all possible a_t we receive the distribution and can calculate the relevant parameters, e.g. mean and variance.

By using Bayes rule to change beliefs it can be shown that the beliefs converge to the true values of the parameters. Hence expectations converge to rational expectations in the narrow sense. This property makes Bayesian learning attractive to use in

[12] If no random variables are present, other models of learning have to be used. In some cases it may be possible to solve the equations directly for the parameters and obtain the true values in a single step. As in most models random variables are incorporated we concentrate on this case here.

[13] For continuous random variables the argument does not change. Instead of probabilities densities have to be used.

rational expectation models. What remains an unsolved problem is how to learn the true structure of the economy.

Appendix D

Mathematical definitions and methods

In this appendix we will provide brief introductions to some mathematical terms and methods which have not been explained in more detail in the main body of the text. For any theorems presented throughout this appendix we will at most give the idea of an proof. Unless otherwise stated, the definitions and theorems are taken from BEEKMANN (1995) and BEEKMANN (1996), where also detailed proofs can be found. Most textbooks on Analysis also provide these proofs, e.g. MARSDEN AND WEINSTEIN (1985).

D.1 Definitions

D.1.1 Closedness and compactness

Let $\mathcal{D} \subseteq \mathbb{R}$ be a set.[1] We then call $a \in \mathcal{D}$ an *accumulation point* if in every environment of a there exists a point $b \neq a$ such that $b \in \mathcal{D}$.

A set $\mathcal{D}$ is called *closed* if every accumulation point of $\mathcal{D}$ is in $\mathcal{D}$. A set $\mathcal{D}$ is *open* if it is not closed.

We call a mapping $\gamma : \mathcal{D} \mapsto \mathcal{D}$ *closed* for all $a \in \mathcal{D}$ with $a^n \to a$, $b^n \in \gamma(a^n)$ and $b^n \to b$, if we have $b \in \gamma(a)$. Geometrically, we can state that the graph of a closed mapping γ can be drawn by a single, uninterrupted line.

[1] In the case that $\mathcal{D} \subseteq \mathbb{R}^N$ with $N > 1$ similar considerations, although less intuitive define a compact set. Instead of real numbers we could use any field. However, as only the real line is used in this work, we do not consider these cases further.

We call a set $\mathcal{D}$ *bounded* if there exist numbers $s, S \in \mathcal{D}$ such that $\forall a \in \mathcal{D} : s \leq a \leq S$.

A set $\mathcal{D}$ is *compact* if it is closed and bounded.

Every closed interval $[a, b] \in \mathbb{R}$ on the real line is compact.

D.1.2 Convexity

We call a set $\mathcal{D} \subseteq \mathbb{R}^N$ *convex* if $\forall a, b \in \mathcal{D}\ \forall \alpha \in [0, 1] : \alpha a + (1 - \alpha)b \in \mathcal{D}$.

Geometrically, a set is convex if the straight line connecting any two points of this set is located entirely in this set. Every interval of the real line, i.e. $[a, b]$,$]a, b]$,$[a, b[$, and $]a, b[$ with $a, b \in \mathbb{R}$, is a convex set.

A function $f : \mathbb{I} \mapsto \mathbb{R}$ is called convex if

$$\forall a, b, x \in \mathbb{I} : a < x < b \Rightarrow f(x) \leq f(a) + \frac{f(b) - f(a)}{b - a}(x - a)$$

.

If f is continuous and twice differentiable, it is convex (concave) if $\forall x \in \mathbb{I} : f''(x) \geq 0$ $(f''(x) \leq 0)$.

A function is convex (concave) if the straight line connecting any two values of f is always above (below) the graph of f.

D.1.3 Metric and Euclidean spaces

Let $\mathcal{D}$ be a set. A *metric* is a mapping $d : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$ with the following properties holding for any $a, b, c \in \mathcal{D}$:

- $d(a, b) > 0$ if $a \neq b$,
- $d(a, a) = 0$,

- $d(a,b) = d(b,a)$, and
- $d(a,c) \leq d(a,b) + d(b,c)$.

We call $(\mathcal{D}, d)$ a *metric space*, i.e. it is a set with a certain metric defined on it.

With $\mathcal{D} = \mathbb{R}^N$ and for any $x \in \mathcal{D}$, where $x = (x_1, x_2, \ldots, x_N)$ the metric $d(a,b) = \sqrt{\sum_{i=1}^{N} |a_k - b_k|^2}$ is called the *Euclidean metric* and the associated space $(\mathcal{D}, d)$ is called a *Euclidean space*.

D.1.4 Hemi-continuity

Define a mapping $\gamma : \mathcal{D} \mapsto \mathcal{E}$, $\mathcal{F} \subseteq \mathcal{E}$, and $\mathcal{C} \subseteq \mathcal{D}$. We call $\gamma(\mathcal{C}) = \bigcup_{a \in \mathcal{F}} \gamma(a)$ the *image* of $\mathcal{F}$ under γ.

The *upper inverse* of $\mathcal{F}$ under γ we define as $\gamma^+[\mathcal{F}] = \{a \in \mathcal{D} | \gamma(a) \subset \mathcal{F}\}$.

We call a mapping *upper hemi-continuous* if the upper inverses of open sets are open.[2]

Let $\mathcal{D} \subseteq \mathbb{R}^n$ and $\mathcal{E} \subseteq \mathbb{R}^m$, where $\mathcal{E}$ is compact. If in this case γ is closed, then γ is upper hemi-continuous.

D.2 Local separation

Let $a \in \mathbb{R}$ and $b \in \mathbb{R}$ with $a \neq b$. Then there exist environments $\mathcal{U}$ and $\mathcal{V}$ of a and b, respectively, such that $\mathcal{U} \cap \mathcal{V} = \emptyset$.

This result is called the *theorem of local separation* or *Hausdorff property*.

[2] This definition is taken from BORDER (1985, p. 55).

D.3 Taylor series expansion

Let $f : \mathcal{D} \mapsto \mathbb{R}$ be $(n+1)$-times differentiable on $\mathcal{D}$ with $\mathcal{D} \subset \mathbb{R}$.[3] We then can define a *Taylor polynome* by

$$T_n : \quad \mathcal{D} \mapsto \mathbb{R} \qquad x \mapsto T_n \equiv \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k, \tag{D.1}$$

where $a \in \mathcal{D}$ and $n \geq 0$ and $f^{(k)}$ denotes the kth derivative of f.

Let there exist an open interval $\mathcal{D}_0 \subset \mathcal{D}$, $a \in \mathcal{D}$ and two constants $c > 0$ and $M > 0$ such that for all $k \geq 0$ and $x \in \mathcal{D}_0$

$$\left|f^{(k)}(x)\right| \leq ck!M^k. \tag{D.2}$$

If this condition is fulfilled, we can expand f into a Taylor series:

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!}(x-a)^k \tag{D.3}$$

for all $x \in \mathcal{D}_0$ with $|x-a| < \frac{1}{M}$.

As high order terms become arbitrarily small, we can approximate f by

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k \tag{D.4}$$

and call this an *expansion of f into a nth order Taylor series around a*. In practice it is always assumed that the conditions for a Taylor series expansion are met.

D.4 Implicit functions

Suppose a system of n not necessarily linear equations with $m+n$ unknown variables, where $m > 0$ and F_i is a mapping $F_i : \mathbb{R}^{m+n} \to \mathbb{R}$.

[3] We can derive similar results for multidimensional functions $f : \mathcal{D} \mapsto \mathbb{R}^M$ with $\mathcal{D} \subset \mathbb{R}^N$.

Define now $F = (F_1, F_2, \ldots, F_n)'$ with $F : \mathcal{M} \to \mathbb{R}$, where $\mathcal{M} \subseteq \mathbb{R}^{n+m}$ and $z = (z_1, z_2, \ldots, z_{n+m}) = (x, y) \in \mathcal{M}$. It is $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Suppose F to be continuously differentiable on $\mathcal{M}$ and that there exists a point $c = (a, b) \in \mathcal{M}$ such that $F(c) = 0$ and $\det F'_y(c) \neq 0$, where F'_y denotes the submatrix of F' containing all partial derivatives of F with respect to $z_{n+1}, \ldots, z_{n+m}$.

With these conditions fulfilled, there exists an open environment $\mathcal{U} \subseteq \mathbb{R}^n$ of a, an open environment $\mathcal{V} \subseteq \mathbb{R}^m$ of b, and a continuously differentiable function $g : \mathcal{U} \mapsto \mathcal{V}$ with the following properties:

- $\mathcal{U} \times \mathcal{V} \equiv \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{m+n} | x \in \mathcal{U}, y \in \mathcal{V} \right\} \subset \mathcal{M}$ and $F(x, g(x)) = 0$ for every $x \in \mathcal{U}$. If $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{U} \times \mathcal{V}$ with $F(x, y) = 0$ then $y = g(x)$.
- $g'(x) = -[F'_y(x, g(x))]^{-1} F'_x(x, g(x))$ for every $x \in \mathcal{U}$, where F'_x denotes the submatrix of F' containing all partial derivatives with respect to $z_1, \ldots, z_n$.

This result is known as the *theorem on implicit functions*.

D.5 Dynamic programming[4]

When individuals have not only to make decisions that are optimal at a given point of time by choosing an optimal value of the *control variables*, u, but that these values have to be chosen optimal over a certain period of time, they have to determine an optimal *time path* for the variables. In many cases an additional problem arises, that the environment determining the outcome of this optimization problem changes through a changing *state variable*, x, which may be influenced by the control variables.

We define a function $I(x, u, t)$ which measures the payoff at a certain point of time, t. The aim of the individual now is to maximize the payoffs he receives over time,

[4] This section is based on INTRILIGATOR (1971).

i.e. the *control problem* is given by[5]

$$\max_{u(t)} J = \int_{t_0}^{t_1} I(x,u,t)dt, \tag{D.5}$$

where t_0 and t_1 denote the starting and end point of the considerations. The state variable changes according to the differential equation

$$\frac{\partial x}{\partial t} = f(x,u,t). \tag{D.6}$$

We define the solution to the control problem by $J^*(x,t)$ and call this the *optimal performance function.*

The *principle of optimality* now requires that regardless of the current state the remaining decisions have to be optimal. Therewith at point $t+\Delta t$ with state $x+\Delta x$ the optimal performance function has to be $J^*(x+\Delta x, t+\Delta t)$. We can now write the optimal performance function as

$$J^*(x,t) = \max_{u(t)}\{I(x,u,t)\Delta t + J^*(x+\Delta x, t+\Delta t)\}, \tag{D.7}$$

which is known as the *fundamental recurrence relation.*[6] Here $I(x,u,t)\Delta t$ denotes the payoff in the interval $]t, t+\Delta t]$. Approximating the second term in brackets by a first order Taylor series around (x,t), we get

$$J^*(x+\Delta x, t+\Delta t) = J^*(x,t) + \frac{\partial J^*}{\partial x}\Delta x + \frac{\partial J^*}{\partial t}\Delta t, \tag{D.8}$$

which gives after inserting into (D.7):

$$0 = \max_{u(t)}\{I(x,u,t)\Delta t + \frac{\partial J^*}{\partial x}\Delta x + \frac{\partial J^*}{\partial t}\Delta t\}. \tag{D.9}$$

Dividing by Δt and taking the limit $\Delta t \to 0$ we get with

$$\lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{\partial x}{\partial t} = f(x,u,t) \tag{D.10}$$

[5] In dynamic programming it is usually assumed that future payoffs are not discounted to their present value.

[6] By taking expectations on both sides, we get the equivalent relation for stochastic variables: $J^*(x,t) = \max_{u(t)}\{E[I(x,u,t)]\Delta t + E[J^*(x+\Delta x, t+\Delta t)]\}$. In the same manner we can include a discount factor for future payoffs by defining $J^*(x+\Delta x, t+\Delta t) = \rho J^*(x+\Delta x, t)$ without changing the argument. In this case the fundamental recurrence relation becomes $J^*(x,t) = \max_{u(t)}\{I(x,u,t)\Delta t + \rho J^*(x+\Delta x, t)\}$.

$$-\frac{\partial J^*}{\partial t} = \max_{u(t)} \left\{ I(u,x,t) + \frac{\partial J^*}{\partial x} f(x,u,t) \right\}. \tag{D.11}$$

This partial differential equation is known as the *Bellman equation*. Solving this equation will give the optimal performance function, given boundary conditions.

D.6 Fixed point theorems[7]

Suppose a set $\mathcal{D}$ and a mapping $\gamma : \mathcal{D} \mapsto \mathcal{D}$. A point $a \in \mathcal{D}$ satisfying $\gamma(a) = a$ is called of *fixed point* of γ. We can define a function $\gamma' = \gamma - \mathrm{id}$ and define a fixed point as any point $a \in \mathcal{D}$ solving $\gamma'(a) = 0$.

Several authors have derived conditions for the existence of fixed point, most widely applied are the versions of BROUWER (1912) and KAKUTANI (1941).

Brouwer's fixed point theorem states that if $\mathcal{D} \subset \mathbb{R}^N$ is convex and compact and $\gamma : \mathcal{D} \mapsto \mathcal{D}$ is continuous, then γ has a fixed point.

As the continuity of mappings is not always fulfilled, this requirement is relaxed in *Kakutani's fixed point theorem*:

If $\mathcal{D} \subset \mathbb{R}^N$ is a compact and convex set and $\gamma : \mathcal{D} \mapsto \mathcal{D}$ is closed or upper-hemicontinuous, where $\gamma(\mathcal{D}) \neq \emptyset$ is convex and compact, then γ has a fixed point.

[7] BORDER (1985) gives an overview of fixed point theorems and their applications on which this section also is based.

Appendix E

Proof of theorem 5

Before we can proceed to proof theorem 5 we have to determine some properties of λ, the cooperative equilibrium and the profits under defection.

To facilitate notation we define $\varphi^C(p,\mu) = \lambda^C(p,\mu)d(p,\mu)$ and $\varphi^D(p,\mu) = \lambda^D(p,\mu)d(p,\mu)$. In order to distinguish between variables and their values in this appendix we will use the denote the derivative of a function $f(x,y)$ with respect to x evaluated at (x_0,y_0) by $\frac{\partial f(x,y)}{\partial x}\Big|_{x=x_0}^{y=y_0}$.

Proposition 1. *λ is continuous in p^i for any μ_i.*

Proof. Suppose that λ is discontinuous at $p^{i\prime}$, which implies that the probability of observing $p^{i\prime}$ is positive. As the distribution of μ_i is continuous by assumption, there has to exist a subset $K \subseteq [\underline{\mu},\overline{\mu}]$ with an infinite and uncountable cardinality,[1] such that for each $\mu_i \in K$ the price $p^{i\prime}$ is chosen.

When $K = [\underline{\mu},\overline{\mu}]$, i.e. the same price is quoted for all signals, $\lambda(p^{i\prime},\mu_i) = \phi$ for all μ_i and all market makers, hence $\frac{\partial \lambda(p^i,\mu)}{\partial p^i}\Big|_{p^i=p^{i\prime}}^{\mu=\mu_i} = 0$. The first order condition for a

[1] The cardinality of a set describes the number of elements in this set.

profit maximum becomes for all $\mu_i \in K$:

$$
\begin{aligned}
(\text{E.1}) \qquad 0 \;=\; & \frac{\partial \lambda\left(p^i,\mu\right)}{\partial p^i}\Big|^{\mu=\mu_i}_{p^i=p^{i\,\prime}} (p^{i\,\prime}-p^*)d(p^{i\,\prime},\mu_i) \\
& +\lambda\left(p^{i\,\prime},\mu_i\right)\left[d(p^{i\,\prime},\mu_i)+(p^{i\,\prime}-p^*)\,\frac{\partial d(p^i,\mu)}{\partial p^i}\Big|^{\mu=\mu_i}_{p^i=p^{i\,\prime}}\right] \\
=\; & \phi\left[d(p^{i\,\prime},\mu_i)+(p^{i\,\prime}-p^*)\,\frac{\partial d(p^i,\mu)}{\partial p^i}\Big|^{\mu=\mu_i}_{p^i=p^{i\,\prime}}\right].
\end{aligned}
$$

This condition is equivalent to the condition for a profit maximum in the cooperative case of the Dutta-Madhavan model. As we have shown in theorem 2 that the optimal price is increasing in the liquidity event, and therewith also in the signal received, condition (E.1) cannot be fulfilled for two different $\mu_i \in K$ at the same $p^{i\,\prime}$. Therefore it is required that K is a strict subset of $[\underline{\mu},\overline{\mu}]$, i.e. there exist signals for which prices $p^i_c \neq p^{i\,\prime}$ are optimal.

If there exist signals for which other prices are optimal, these prices cannot be located in an area above $p^{i\,\prime}$. Because λ is decreasing in p^i, a discontinuity in $p^{i\,\prime}$ implies a jump downwards, hence the expected profits also jump downwards at $p^{i\,\prime}$. Then, with the theorem of local separation[2], there exist areas $I =]p^{i\,\prime}, p^{i\,\prime}+\delta[$ with $\delta > 0$ and $J =]p^{i\,\prime}-\eta, p^{i\,\prime}[$ with $\eta > 0$ in which for all market makers with $\mu_i \notin K$ the expected profits from quoting a price $p^i \in I$ are smaller than from quoting a price $p^i \in J$. Therewith it cannot be optimal to quote a price $p^i \in I$ for any $\mu_i \notin K$ and no market maker will quote a price $p^i \in I$ (those with a signal $\mu_i \in K$ quote $p^{i\,\prime} \notin I$). Figure E.1 illustrates this result.

For this reason λ will be non-decreasing in p^i for any $p^i \in I$, hence there exists a price $p^i \in I$ for any market maker receiving a signal $\mu_i \in K$ such that the expected profits are larger than from quoting $p^{i\,\prime}$.[3] Dealers receiving a signal $\mu_i \in K$ would therefore be better off when quoting a higher price $p^i \in I$. Hence it would be no

[2] See Appendix D.2.

[3] The first order condition for a profit maximum is

$$
\frac{\partial \lambda\left(p^i,\mu_i\right)}{\partial p^i}(p^i-p^*)d(p^i,\mu_i)+\lambda\left(p^i,\mu_i\right)\left[d(p^i,\mu_i)+(p^i-p^*)\frac{\partial d(p^i,\mu_i)}{\partial p^i}\right]=0
$$

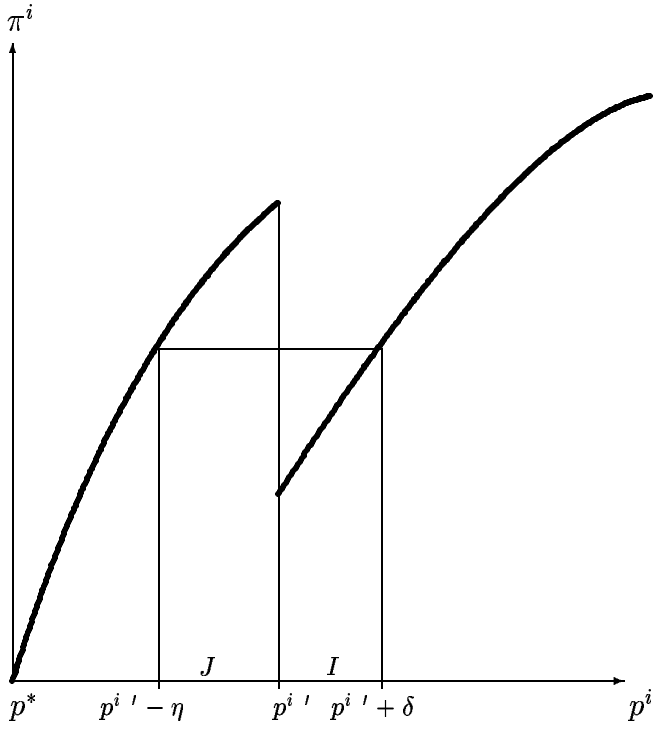


Figure E.1: Profits with a discontinuity of λ

equilibrium to quote $p^{i\prime}$ for any $\mu_i \in K$. This holds as long as the cardinality of K is infinite and uncountable. Only if the cardinality of K is countable, i.e. only for isolated signals, rather than intervals of μ_i, it is optimal to quote p^i for any $\mu_i \in K$. If the cardinality of K is countable, however, λ is continuous. Therefore in equilibrium λ is continuous. □

Proposition 2. $\frac{\partial\lambda(p^i,\mu_i)}{\partial p^i}$ *is upper hemi-continuous in* p^i *for any* μ_i.

Proof. By proposition 1 λ is continuous, hence if λ is not differentiable at $p^{i\prime}$, the function has a kink at $p^{i\prime}$ and the derivative is discontinuous. Therewith at $p^{i\prime}$ the derivative of λ with respect to p^i can take any values between $\lim_{\theta\to 0} \left.\frac{\partial\lambda(p^i,\mu)}{\partial p^i}\right|_{p^i=p^{i\prime}+\theta}^{\mu=\mu_i}$

for $p^i \notin I$ and

$$\lambda\left(p^i,\mu_i\right)\left[d(p^i,\mu_i) + (p^i - p^*)\frac{\partial d(p^i,\mu_i)}{\partial p^i}\right] = 0$$

for $p^i \in I$ as λ is non-decreasing in p^i. Suppose that the expected profits are decreasing for $p^{i\prime}$, hence the second relation is negative, then also the first relation is negative in $p^{i\prime}$ and it would be optimal to quote a lower price than $p^{i\prime}$ and $p^{i\prime}$ cannot be an equilibrium.

and $\lim_{\theta\to 0} \frac{\partial\lambda\left(p^i,\mu\right)}{\partial p^i}\Big|_{p^i=p^{i\,\prime}-\theta}^{\mu=\mu_i}$, what implies $\frac{\partial\lambda\left(p^i,\mu\right)}{\partial p^i}$ to be closed by the definition in appendix D.1.1.

With the same argumentation with which we showed that λ is continuous, we can show that its slope has to be bounded,[4] hence $\frac{\partial\lambda\left(p^i,\mu_i\right)}{\partial p^i}$ is bounded. Therewith the values of $\frac{\partial\lambda\left(p^i,\mu_i\right)}{\partial p^i}$ are on the real line and we can define an interval $I = [\underline{\lambda}, 0]$ such that $\frac{\partial\lambda\left(p^i,\mu_i\right)}{\partial p^i} \in I$ for all $p^i \in \mathcal{S}^i$ and $\mu_i \in [\underline{\mu}, \overline{\mu}]$, i.e. the values of $\frac{\partial\lambda\left(p^i,\mu_i\right)}{\partial p^i}$ are in a closed set. This shows the values of $\frac{\partial\lambda\left(p^i,\mu_i\right)}{\partial p^i}$ to be in a compact set.

From the criteria for upper hemi-continuity in appendix D.1.4, we have shown that $\frac{\partial\lambda\left(p^i,\mu_i\right)}{\partial p^i}$ is upper hemi-continuous. □

Proposition 3. *The price and expected profits of any equilibrium are increasing in* μ_i.

Proof. Assume that for some $\mu_i' > \mu_i$ we find $p^i(\mu_i') < p^i(\mu_i)$. Then it would be suboptimal to quote $p^i(\mu_i)$ when receiving the signal μ_i' and to quote $p^i(\mu_i')$ when receiving μ_i:

$$\begin{aligned}
\lambda(p^i(\mu_i'),\mu_i')\left(p^i(\mu_i')-p^*\right)d\left(p^i(\mu_i'),\mu_i'\right) &> \lambda(p^i(\mu_i),\mu_i')\left(p^i(\mu_i)-p^*\right)d\left(p^i(\mu_i),\mu_i'\right),\\
\lambda(p^i(\mu_i),\mu_i)\left(p^i(\mu_i)-p^*\right)d\left(p^i(\mu_i),\mu_i\right) &> \lambda(p^i(\mu_i'),\mu_i)\left(p^i(\mu_i')-p^*\right)d\left(p^i(\mu_i'),\mu_i\right).
\end{aligned}$$

Combining these inequalities gives

$$\frac{\lambda(p^i(\mu_i),\mu_i)}{\lambda(p^i(\mu_i'),\mu_i)}\frac{d(p^i(\mu_i),\mu_i)}{d(p^i(\mu_i'),\mu_i)} > \frac{\lambda(p^i(\mu_i),\mu_i')}{\lambda(p^i(\mu_i'),\mu_i')}\frac{d(p^i(\mu_i),\mu_i')}{d(p^i(\mu_i'),\mu_i')}.$$

This result contradicts the assumption of the monotone demand ratio. Hence the above relation cannot be fulfilled and the optimal price cannot decrease in μ_i.

From the assumption of monotonicity of the expected demand in μ_i, expected profits must be increasing in μ_i when holding prices constant. If by increasing prices, expected profits would decrease, this strategy is not optimal and lower prices would be quoted. Therewith expected profits also increase in μ_i. □

[4] If the slope of λ would not be bounded at $p^{i\,\prime}$, it must be $-\infty$, i.e. the graph would be a vertical line at $p^{i\,\prime}$ and we could again define environments I and J such that in analogy to proposition 1 this situation could be ruled out.

Proposition 4. *The profit maximizing price of a market maker is bounded.*

Proof. The profit maximizing price, i.e. the price maximizing (5.42), denoted p_c^i, will never be such that the expected profits are negative, hence $p_c^i \geq p^*$. The results from proposition 3 that any equilibrium price is increasing in μ_i, implies that, as μ_i is bounded by assumption, the maximal price that can be achieved is $p_{max}^i = p_c^i(\overline{\mu})$. The concavity of profits further implies that $p_c^i(\overline{\mu}) < \infty$. Therewith $p^* \leq p_c^i \leq p_{max}^i < \infty$ and the profit maximizing price is bounded. □

Proposition 5. *A profit maximizing equilibrium exists.*

Proof. As p_c^i is bounded it is in a compact set and as $p^i \in \mathbb{R}$ this set is also convex.[5] By assumption $d(p^i, \mu_i)$ and $\frac{\partial d(p^i,\mu_i)}{\partial p^i}$ are continuous, from proposition 1 $\lambda(p^i, \mu_i)$ is continuous and proposition 2 shows $\frac{\partial \lambda(p^i,\mu_i)}{\partial p^i}$ to be upper hemi-continuous.

In the profit maximizing equilibrium market makers maximize their present value of future expected profits, which is equivalent to maximizing expected profits in each period of time. The first order condition for a maximum is

$$\text{(E.2)} \qquad \frac{\partial \pi(p^i, \mu_i)}{\partial p^i} = \frac{\partial \lambda\,(p^i, \mu_i)}{\partial p^i}(p^i - p^*)d(p^i, \mu_i) + \lambda\left(p^i, \mu_i\right)\left[d(p^i, \mu_i) + (p^i - p^*)\frac{\partial d(p^i, \mu_i)}{\partial p^i}\right] = 0.$$

With the foresaid $\frac{\partial \pi(p^i,\mu_i)}{\partial p^i}$ is upper hemi-continuous. As the cooperative price has been shown to be bounded in proposition 4, we have a fixed point according to Kakutani's fixed point theorem, i.e. there exists a p^i fulfilling (E.2).[6] At least one of these fixed points has to be a profit maximum and hence an equilibrium, because $\pi(p^*, \mu_i) = \pi(\infty, \mu_i) = 0$ and for all μ_i the expected profits are assumed to be concave. □

[5] Any bounded and compact subset of $\mathbb{R}$ is convex.

[6] Fixed point theorems are briefly considered in Appendix D.6.

For simplicity in the remainder we will call this profit maximizing equilibrium the cooperative equilibrium, the associated profits of this equilibrium the cooperative profits and the corresponding prices the cooperative prices.

We will now characterize the optimal defection of a market maker not willing to cooperate. It will be necessary to distinguish between the distribution of prices under cooperation and defection, so that they will be denoted λ^C and λ^D, respectively. The optimal defective (cooperative) price will be denoted $p_D^{i,*}$ ($p_c^{i,*}$), the variables are denoted p_D^i and p_c^i, respectively and the profit function under cooperation and defection π^C and π^D, respectively.

Proposition 6. *The optimal defective price is strictly below the cooperative price.*

Proof. As upon defection the market maker also maximizes his expected profits, we see the first order conditions for the optimal prices $p_c^{i,*}$ and $p_D^{i,*}$ to be

$$\begin{aligned}(\text{E.3}) \qquad 0 \;=\; & \left.\frac{\partial \pi^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} = \left.\frac{\partial \lambda^C\left(p_c^i,\mu\right)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} (p_c^i - p^*) d(p_c^i,\mu_i) \\ & + \lambda^C\left(p_c^i,\mu_i\right)\left[d(p_c^i,\mu_i) + (p_c^i - p^*)\left.\frac{\partial d(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}\right],\end{aligned}$$

$$\begin{aligned}(\text{E.4}) \qquad 0 \;=\; & \left.\frac{\partial \pi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} = \left.\frac{\partial \lambda^D\left(p_D^i,\mu\right)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} (p_D^i - p^*) d(p_D^i,\mu_i) \\ & + \lambda^D\left(p_D^i,\mu_i\right)\left[d(p_D^i,\mu_i) + (p_D^i - p^*)\left.\frac{\partial d(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}\right].\end{aligned}$$

If the price a market maker quotes in the cooperative equilibrium changes, he expects the other market makers also to change their prices into the same direction. We therefore can state that the probability of receiving a trade does only change slightly. Upon defection he expects the other market makers not to change their quotes, hence he can expect to influence the probability of receiving a trade much more:

$$(\text{E.5}) \qquad \left.\frac{\partial \lambda^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} < \left.\frac{\partial \lambda^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} < 0.$$

Suppose now that the prices with defection and cooperation differ only marginally as in the case of perfect knowledge of the liquidity event. In this case the second

terms in (E.3) and (E.4) are identical. Due to (E.5) the first term in (E.4) is smaller than in (E.3). Therewith, if (E.3) is fulfilled, the second equation is negative. The concavity of the profit function then implies the optimal defective price to be strictly below the cooperative price. □

As a last step before proofing theorem 5, we will compare the profits from defection and cooperation.

Proposition 7. *For any liquidity event μ_i and any prices p_c^i and p_D^i it is* $0 \leq \frac{\partial \pi^C(p_c^i,\mu)}{\partial \pi^D(p_D^i,\mu)}\Big|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} < 1$, *i.e. the defective profits react much more sensitive to any change in the variables than the cooperative profits.*

Proof. We will use the above claim in the form

$$\frac{\partial \pi^C(p_c^i,\mu)}{\partial \pi^D(p_D^i,\mu)}\Big|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} = \frac{\frac{\partial \pi^C(p_c^i,\mu)}{\partial p_c^i}\Big|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}}{\frac{\pi^D(p_D^i,\mu)}{\partial p_c^i}\Big|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}} < 1, \tag{E.6}$$

or equivalently,

$$\frac{\partial \pi^C(p_c^i,\mu)}{\partial p_c^i}\Big|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} < \frac{\pi^D(p_D^i,\mu)}{\partial p_c^i}\Big|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}. \tag{E.7}$$

The respective derivatives are given by

$$\begin{aligned} \frac{\partial \pi^C(p_c^i,\mu)}{\partial p_c^i}\Big|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} &= \\ &= \frac{\partial \lambda^C(p_c^i,\mu)}{\partial p_c^i}\Big|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} (p_c^i - p^*) d(p_c^i,\mu_i) \\ &+ \lambda^C(p_c^i,\mu_i)\left[d(p_c^i,\mu_i) + (p_c^i - p^*)\frac{\partial d(p_c^i,\mu)}{\partial p_c^i}\Big|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}\right], \end{aligned} \tag{E.8}$$

$$
\begin{aligned}
(\text{E.9}) \quad & \left.\frac{\partial \pi^D(p_D^i,\mu)}{\partial p_c^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} = \\
& = \left.\frac{\partial \lambda^D\left(p_D^i,\mu\right)}{\partial p_c^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} (p_D^i - p^*) d(p_D^i,\mu_i) \\
& + \lambda^D\left(p_D^i,\mu_i\right)\left[d(p_D^i,\mu_i)\left.\frac{\partial p_D^i}{\partial p_c^i}\right|_{p_D^i=p_D^{i,*}} + (p_D^i - p^*)\left.\frac{\partial d(p_D^i,\mu)}{\partial p_c^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}\right] \\
& = \left.\frac{\partial p_D^i}{\partial p_c^i}\right|_{p_D^i=p_D^{i,*}} \left\{ \left.\frac{\partial \lambda^D\left(p_D^i,\mu\right)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} (p_D^i - p^*) d(p_D^i,\mu_i) \right. \\
& \left. + \lambda^D\left(p_D^i,\mu_i\right)\left[d(p_D^i,\mu_i) + (p_D^i - p^*)\left.\frac{\partial d(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}\right]\right\}.
\end{aligned}
$$

We will compare the terms in these expressions to proof the claim. At first let us compare the term in curled brackets from (E.9) with (E.8).

By using first order Taylor series approximations we see that

$$
\begin{aligned}
(\text{E.10}) \qquad \lambda^D(p_D^{i,*},\mu_i) - \lambda^D(p^*,\mu_i) &= \left.\frac{\partial \lambda^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} (p_D^i - p^*), \\
\lambda^C(p_c^{i,*},\mu_i) - \lambda^C(p^*,\mu_i) &= \left.\frac{\partial \lambda^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} (p_c^i - p^*).
\end{aligned}
$$

Obviously as $p_c^i > p_D^i$ we find that $\lambda^D(p_D^{i,*},\mu_i) > \lambda(p_c^{i,*},\mu_i)$. If the collusive price is at the reservation price of the asset, then it is obvious that also the defective price is at this level, because by quoting a lower price he would make a loss, hence all market makers would then quote the reservation price. If on the other hand the defective price is the reservation price, then a defective market maker would make no profits. As the profits in the current period for a market maker colluding cannot be higher, he also has to receive zero profits, hence market makers colluding will also quote their reservation price.[7] therewith in both cases all market makers quote the same price, their reservation price, and we can reasonably assume all market makers to have the same probability of receiving the order in each case and hence the two expressions are equal and therewith the first expression is larger than the second if we take into account that from the monotonicity of the demand function $d(p_D^{i,*},\mu_i) > d(p_c^{i,*},\mu_i)$.

[7] As p_c^i is bounded from proposition 4, we find that only for $p_c^i = p^*$ profits can become zero.

In the same manner we derive with a first order Taylor series approximation that

$$\begin{aligned} d(p_D^{i,*},\mu_i) - d(p^*,\mu_i) &= \left.\frac{\partial d(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} (p_D^i - p^*), \\ d(p_c^{i,*},\mu_i) - d(p^*,\mu_i) &= \left.\frac{\partial d(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} (p_c^i - p^*). \end{aligned} \tag{E.11}$$

With $p_D^i < p_c^i$ from proposition 6 and the continuity of the demand function, it is obvious that the first expression is larger.

As also obviously $d(p_D^{i,*},\mu_i) > d(p_c^{i,*},\mu_i)$ and $\lambda^D(p_D^{i,*},\mu_i) > \lambda(p_c^{i,*},\mu_i)$ because of $p_c^i > p_D^i$, we see that also the second summand in (E.8) is smaller than in (E.9).

therewith we find that the entire expression in (E.8) is smaller than that in curled brackets of (E.9). What now remains to be proved is $\frac{\partial p_D^i}{\partial p_c^i} > 1$.

The first derivatives of the expected profits for a defective and collusive behavior are according to (E.3) and (E.4) given by:

$$F(\mu_i, p_D^i, p_c^i) = \begin{bmatrix} \frac{\partial \varphi^D(p_D^i,\mu_i)}{\partial p_D^i}(p_D^i - p^*) + \varphi^D(p_D^i,\mu_i) \\ \frac{\partial \varphi^C(p_c^i,\mu_i)}{\partial p_c^i}(p_c^i - p^*) + \varphi^C(p_c^i,\mu_i) \end{bmatrix}. \tag{E.12}$$

The optimal prices are defined as

$$\begin{bmatrix} p_D^{i,*} \\ p_c^{i,*} \end{bmatrix} = \left[\begin{bmatrix} p_D^i \\ p_c^i \end{bmatrix} \middle| F(p_D^i, p_c^i, \mu_i) = 0 \right]. \tag{E.13}$$

The Jacobian of F is

$$F'(\mu_i, p_D^i, p_c^i) = [F_L \quad F_p], \tag{E.14}$$

where

$$F_L = \begin{bmatrix} F_L^1 \\ F_L^2 \end{bmatrix} \tag{E.15}$$

with

$$\begin{aligned} F_L^1 &= \left.\frac{\partial^2 \varphi^D(p_D^i,\mu)}{\partial p_D^i \partial \mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} (p_D^i - p^*) \\ &\quad + \left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} \left.\frac{\partial p_D^i(\mu)}{\partial \mu}\right|_{\mu=\mu_i} \\ &\quad + \left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\right|_{\mu=\mu_i}^{p_D^i=p_D^{i,*}}, \end{aligned} \tag{E.16}$$

$$F_L^2 = \left.\frac{\partial^2\varphi^C(p_c^i,\mu)}{\partial p_c^i \partial\mu}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}(p_c^i-p^*) + \left.\frac{\partial\varphi^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}\left.\frac{\partial p_c^i(\mu)}{\partial\mu}\right|_{\mu=\mu_i} + \left.\frac{\partial\varphi^C(p_c^i,\mu)}{\partial\mu}\right|_{\mu=\mu_i}^{p_c^i=p_c^{i,*}}. \tag{E.17}$$

$$F_p = \begin{bmatrix} F_p^1 & F_p^2 \\ F_p^2 & F_p^3 \end{bmatrix}, \tag{E.18}$$

with

$$\begin{aligned} F_p^1 &= \left.\frac{\partial^2\varphi^D(p_D^i,\mu)}{\partial {p_D^i}^2}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p_D^i-p^*) + 2\left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}, \\ F_p^2 &= \left.\frac{\partial^2\varphi^D(p_D^i,\mu)}{\partial p_D^i \partial p_c^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p_D^i-p^*) + 2\left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial p_c^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} \\ &= \left.\frac{\partial^2\varphi^C(p_c^i,\mu)}{\partial p_c^i \partial p_D^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}(p_c^i-p^*) + 2\left.\frac{\partial\varphi^C(p_c^i,\mu)}{\partial p_D^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}, \\ F_p^3 &= \left.\frac{\partial^2\varphi^C(p_c^i,\mu)}{\partial {p_c^i}^2}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}(p_c^i-p^*) + 2\left.\frac{\partial\varphi^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}. \end{aligned} \tag{E.19}$$

We further obtain

$$F_p^{-1} = \frac{1}{F_p^1 F_p^3 - (F_p^2)^2}\begin{bmatrix} F_p^3 & -F_p^2 \\ -F_p^2 & F_p^1 \end{bmatrix}. \tag{E.20}$$

Using first and second order Taylor series approximations we get the following relations:

$$\varphi^D(p^*,\mu_i) = \varphi^D(p_D^{i,*},\mu_i) + \left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p^*-p_D^i) + \frac{1}{2}\left.\frac{\partial^2\varphi^D(p_D^i,\mu)}{\partial {p_D^i}^2}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p^*-p_D^i)^2,$$

what with $p^* - p_D^i \leq 0$ becomes

$$\left.\frac{\partial^2\varphi^D(p_D^i,\mu)}{\partial {p_D^i}^2}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p^*-p_D^i) = -2\frac{\varphi^D(p^*,\mu_i)-\varphi^D(p_D^{i,*},\mu_i)}{p_D^i-p^*} - 2\left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}.$$

$$\varphi^D(p^*,\mu_i) = \varphi^D(p_D^{i,*},\mu_i) + \left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p^* - p_D^i),$$

which can be solved for

$$\frac{\varphi^D(p^*,\mu_i) - \varphi^D(p_D^{i,*},\mu_i)}{p_D^i - p^*} = -\left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p^*}^{\mu=\mu_i}.$$

Inserting these relations we see that

$$F_p^1 = 2\left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p^*}^{\mu=\mu_i}. \tag{E.21}$$

Applying the same steps as before we receive

$$F_p^3 = 2\left.\frac{\partial\varphi^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p^*}^{\mu=\mu_i}. \tag{E.22}$$

The same method gives us

$$\left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial\mu}\right|_{p_D^i=p^*}^{\mu=\mu_i} = \left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial\mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} + \left.\frac{\partial^2\varphi^D(p_D^i,\mu)}{\partial\mu\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p^* - p_D^i),$$

which transforms into

$$\left.\frac{\partial^2\varphi^D(p_D^i,\mu)}{\partial\mu\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}(p_D^i - p^*) = \left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial\mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} - \left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial\mu}\right|_{p_D^i=p^*}^{\mu=\mu_i}.$$

Inserting this relation results in

$$\begin{aligned} F_L^1 \;=\; & 2\left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial\mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} - \left.\frac{\partial\varphi^D(p_D^i,\mu)}{\partial\mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} \\ & + \left.\frac{\partial\varphi(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}\left.\frac{\partial p_D^i(\mu)}{\partial\mu}\right|_{\mu=\mu_i}, \end{aligned} \tag{E.23}$$

and in the same manner we derive

$$\begin{aligned} F_L^2 \;=\; & 2\left.\frac{\partial\varphi^C(p_c^i,\mu)}{\partial\mu}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} - \left.\frac{\partial\varphi^C(p_c^i,\mu)}{\partial\mu}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} \\ & + \left.\frac{\partial\varphi(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}\left.\frac{\partial p_c^i(\mu)}{\partial\mu}\right|_{\mu=\mu_i}, \end{aligned} \tag{E.24}$$

Let us for simplicity assume that the cross-derivative of expected profits, i.e. F_p^2, is sufficiently close to zero such that it can be neglected in the further analysis. We

can show at the cost of additional calculations, using the same methods as before, that the results do not change without making this assumption.[8]

As the diagonal elements of (E.18) are the second derivatives of the expected profits, they both have to be negative in a maximum, therewith the determinant of F_p has to be positive in equilibrium. This enables us to apply the theorem on implicit functions to derive a solution for our problem. With inserting the expressions for F_p^1, F_p^3, F_L^1 and F_L^2 we get

$$\text{(E.25)} \qquad \begin{bmatrix} \frac{\partial p_D^i(\mu)}{\partial \mu}\Big|_{\mu=\mu_i} \\ \frac{\partial p_c^i(\mu)}{\partial \mu}\Big|_{\mu=\mu_i} \end{bmatrix} = -F_p^{-1}F_L = -\begin{bmatrix} \frac{F_L^1}{F_p^1} \\ \frac{F_L^2}{F_p^3} \end{bmatrix}.$$

Inserting these expression from the relations above and solving the corresponding equations for $\frac{\partial p_D^i(\mu)}{\partial \mu}\Big|_{\mu=\mu_i}$ and $\frac{\partial p_c^i(\mu)}{\partial \mu}\Big|_{\mu=\mu_i}$, respectively, gives us

$$\text{(E.26)} \qquad \frac{\partial p_D^i(\mu)}{\partial \mu}\Big|_{\mu=\mu_i} = \frac{2\,\frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\Big|^{\mu=\mu_i}_{p_D^i=p_D^{i,*}} - \frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\Big|^{\mu=\mu_i}_{p_D^i=p^*}}{-2\,\frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\Big|^{\mu=\mu_i}_{p_D^i=p^*} - \frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\Big|^{\mu=\mu_i}_{p_D^i=p_D^{i,*}}},$$

$$\text{(E.27)} \qquad \frac{\partial p_c^i(\mu)}{\partial \mu}\Big|_{\mu=\mu_i} = \frac{2\,\frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\Big|^{\mu=\mu_i}_{p_c^i=p_c^{i,*}} - \frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\Big|^{\mu=\mu_i}_{p_c^i=p^*}}{-2\,\frac{\partial \varphi^C(p_c^i,\mu)}{\partial p_c^i}\Big|^{\mu=\mu_i}_{p_c^i=p^*} - \frac{\partial \varphi^C(p_c^i,\mu)}{\partial p_c^i}\Big|^{\mu=\mu_i}_{p_c^i=p_c^{i,*}}}.$$

Using first order Taylor series approximations we get

$$\text{(E.28)} \qquad \varphi^D(p^*,\mu_i) = \varphi^D(p_D^{i,*},\mu_i) + \frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\Big|^{\mu=\mu_i}_{p_D^i=p_D^{i,*}}(p^* - p_D^{i,*}),$$

$$\text{(E.29)} \qquad \varphi^D(p_D^{i,*},\mu_i) = \varphi^D(p^*,\mu_i) + \frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\Big|^{\mu=\mu_i}_{p_D^i=p^*}(p_D^{i,*} - p^*).$$

When solving for $\frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\Big|^{\mu=\mu_i}_{p_D^i=p^*}$ we see that

$$\text{(E.30)} \qquad \frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\Big|^{\mu=\mu_i}_{p_D^i=p_D^{i,*}} = \frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\Big|^{\mu=\mu_i}_{p_D^i=p^*}.$$

[8] This assumption is reasonable as the marginal profits in case of choosing the cooperation (defection) are likely to depend not much on the prices in the not chosen, and therefore hypothetical case of defection (cooperation).

A similar result is obtained for the cooperative case. Hence (E.26) and (E.27) become

$$\text{(E.31)} \qquad \left.\frac{\partial p_D^i(\mu)}{\partial \mu}\right|_{\mu=\mu_i} = -\frac{1}{3}\frac{2\left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} - \left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\right|_{p_D^i=p^*}^{\mu=\mu_i}}{\left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p^*}^{\mu=\mu_i}},$$

$$\text{(E.32)} \qquad \left.\frac{\partial p_c^i(\mu)}{\partial \mu}\right|_{\mu=\mu_i} = -\frac{1}{3}\frac{2\left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} - \left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_c^i=p^*}^{\mu=\mu_i}}{\left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p^*}^{\mu=\mu_i}}.$$

We can reasonably suppose that if all market makers apply the same price in collusion as in defection, p^*, that the demand they face, $\lambda^D(p^*,\mu_i)d(p^*,\mu_i)$ and $\lambda^C(p^*,\mu_i)d(p^*,\mu_i)$ respectively, reacts to a change in the price applied and the signal to the same degree:

$$\text{(E.33)} \qquad \left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial p_c^i}\right|_{p_c^i=p^*}^{\mu=\mu_i} = \left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial p_D^i}\right|_{p_D^i=p^*}^{\mu=\mu_i},$$

$$\text{(E.34)} \qquad \left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_c^i=p^*}^{\mu=\mu_i} = \left.\frac{\partial \varphi^D(p_D^i,\mu_i)}{\partial \mu}\right|_{p_D^i=p^*}^{\mu=\mu_i}.$$

Using these relations we see that

$$\text{(E.35)} \qquad \frac{\partial p_D^i}{\partial p_c^i} = \frac{\left.\frac{\partial p_D^i(\mu)}{\partial \mu}\right|_{\mu=\mu_i}}{\left.\frac{\partial p_c^i}{\partial \mu_i}\right|_{\mu=\mu_i}} = \frac{2\left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} - \left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\right|_{p_D^i=p^*}^{\mu=\mu_i}}{2\left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} - \left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_c^i=p^*}^{\mu=\mu_i}}.$$

The concavity of the expected demand in μ_i implies that with $p_D^i < p_c^i$

$$\text{(E.36)} \qquad \left.\frac{\partial \varphi^D(p_D^i,\mu)}{\partial \mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} > \left.\frac{\partial \varphi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} > 0.$$

As the last term in the numerator and denominator is equal in both cases, we see that the numerator exceeds the denominator and hence we have shown that $\frac{\partial p_D^i}{\partial p_c^i} > 1$.

We therewith have proved our claim that $\left.\frac{\partial \pi^C(p_c^i,\mu)}{\partial \pi^D(p_D^i,\mu)}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} < 1$.

Using the result from proposition 3 that any equilibrium price increases in μ_i, i.e. $\left.\frac{\partial \pi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} \geq 0$ and $\left.\frac{\partial \pi^D(p_D^i,\mu)}{\partial \mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i} \geq 0$, we can immediately derive that $\left.\frac{\partial \pi^C(p_c^i,\mu)}{\partial \pi^D(p_D^i,\mu)}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i} = \frac{\left.\frac{\partial \pi^C(p_c^i,\mu)}{\partial \mu}\right|_{p_c^i=p_c^{i,*}}^{\mu=\mu_i}}{\left.\frac{\partial \pi^D(p_D^i,\mu)}{\partial \mu}\right|_{p_D^i=p_D^{i,*}}^{\mu=\mu_i}} \geq 0$, what completes the proof. □

We now have all elements to give a proof of theorem 5, which we repeat here for convenience:

Theorem 5. *If* $\forall i \in \mathcal{P} : \rho > \rho_0^i = 1 - \frac{\pi(p_c^i, \mu_i)}{E[J(\mu_i)]}$ *a collusive equilibrium exists with the following properties:*

There exist $\underline{\mu} \leq \mu^1 \leq \mu_c \leq \mu^2 \leq \overline{\mu}$ *such that for*

- $\mu_i < \mu^1$ *the collusive price equals the cooperative price and is increasing in* μ_i,
- $\mu^1 \leq \mu_i \leq \mu^2$ *the collusive price remains constant and for*
- $\mu_i > \mu^2$ *the collusive price is decreasing in* μ_i.

If $\exists i \in \mathcal{P} : \rho < \rho_0^i$ *the only equilibrium is to quote competitive prices.*

Proof. From constraint (5.44),

$$\pi^D(p_D^i, \mu_i) - \pi^C(p_c^i, \mu_i) \leq \rho E\left[J(\mu_i)\right],$$

we see that the right hand side is a constant and proposition 7 suggests that the left hand side is increasing in μ_i. As all market makers are alike, except for the signal they receive, they will receive the same profits for any given signal. That is why there will exist a common signal μ_c for all market makers such that this constraint becomes an equality for the cooperative price. If for $\mu_i = \underline{\mu}$ this constraint is violated, define $\mu_c = \underline{\mu}$ and if it is still fulfilled for $\overline{\mu}$, define $\mu_c = \overline{\mu}$. For $\mu_i < \mu_c$ the price that can be quoted without fearing a defection, is the cooperative price, which from proposition 3 is increasing in μ_i.

For a signal $\mu_i > \mu_c$ quoting the cooperative price would violate the constraint, hence to avoid a defection a lower price has to be quoted, reducing the left hand side according to proposition 7.

However, at $\mu_i = \mu_c$ the quoted price would be the largest of all possible prices. Hence all market makers receiving another signal will quote a lower price as they

are alike in all remaining aspects and hence for any given signal quote the same price. With the continuity of the distribution function, the probability of another market maker receiving the same signal is zero. Hence, the probability of other market makers quoting a lower price is one and the probability of receiving the order flow is zero and so are expected profits.

It would be a more profitable strategy to reduce the price and increase the probability of receiving the order flow and therewith increase expected profits. But by only lowering the price for $\mu_i = \mu_c$, the prices in the nearest neighborhood would be the highest prices, so that they will also be lowered. They will only be lowered to the same price as the price quoted for $\mu_i = \mu_c$, otherwise for $\mu_i = \mu_c$ the price would be the highest again. Therefore in an area around μ_c the prices will be equal for any μ_i, hence there exist $\mu^1 \leq \mu_c \leq \mu^2$ in which prices are equal. This area has not necessarily to be symmetric around μ_c and obviously the more dispersed signals are, i.e. the less precise the signal is, the larger we can expect this area to be.

As a cooperative equilibrium exists due to proposition 5, also an equilibrium exists if a restriction is applied in determining the equilibrium.[9] It only has to be shown under which conditions this constraint is fulfilled. The expected future profits are according to (5.42) and with using (5.43) given by

$$\text{(E.37)} \qquad J(\mu_i) \begin{cases} = \pi^C(p_c^i, \mu_i) + \rho E\left[J(\mu_i)\right] & \text{for } \mu_i < \mu^1 \\ \leq \pi^C(p_c^i, \mu_i) + \rho E\left[J(\mu_i)\right] & \text{for } \mu^1 \leq \mu_i \leq \mu_c \\ \leq \pi^D(p_D^i, \mu_i) & \text{for } \mu_c < \mu_i \leq \mu^2 \\ = \pi^D(p_D^i, \mu_i) & \text{for } \mu_i > \mu^2 \end{cases} .$$

From (5.43) we furthermore see that

$$\text{(E.38)} \qquad J(\mu_i) \leq \pi^D(p_D^i, \mu_i) \leq \pi^C(p_c^i, \mu_i) + \rho E\left[J(\mu_i)\right],$$

taking expectations on both sides this becomes

$$\text{(E.39)} \qquad E\left[J(\mu_i)\right] \leq \pi(p_c^i, \mu_i) + \rho E\left[J(\mu_i)\right].$$

[9] This result can be derived from the maximum theorem in conjunction with Kakutani's fixed point theorem, see e.g. BORDER (1985, pp. 63 ff.).

Solving for ρ gives the condition for a collusive equilibrium to exist:[10]

$$\rho \geq \rho_0^i = 1 - \frac{\pi(p_c^i, \mu_i)}{E\left[J(\mu_i)\right]}, \tag{E.40}$$

i.e. for market makers being sufficiently patient, constraint (5.43) is fulfilled and a collusive equilibrium exists, otherwise (5.43) will always be violated and competitive pricing will be applied.[11] □

It has to noted that because for all $\mu^1 \leq \mu_i \leq \mu^2$ the optimal price will be the same, the probability of observing the associated price $p^i(\mu_i)$, $\mu_i \in [\mu^1, \mu^2]$, is strictly positive and hence λ is discontinuous at $p^i(\mu_i)$. However, this does not contradict proposition 1 as this discontinuity arises from the constraint applied to the maximization, while in proposition 1 we investigated the unconstrained maximization.

With the same arguments also the result presented in the main text that for small signals competitive pricing may be applied and hence to observe the competitive price p^* has a positive probability and therewith λ is discontinuous at p^*, does also not violate proposition 1.

[10] It is easy to show that $E\left[J(\mu_i)\right] > 0$ for the above described pricing rule. For $E\left[J(\mu_i)\right] = 0$, i.e. competitive pricing, this relation is always fulfilled as $\pi(p_c^i, \mu_i) = 0$.

[11] The solution in Dutta and Madhavan (1997) that $\rho \geq 1 - \phi$ arises here as a special case if $\pi(p_c^i, \mu_i) = \phi E\left[J(\mu_i)\right]$.

Bibliography

ABREU, D. (1988): Towards a Theory of Discounted Repeated Games. In: Econometrica, 56, 383–396.

ACKERT, L. F. AND CHURCH, B. K. (1999): Bid-Ask Spreads in Multiple Dealer Settings: Some Experimental Evidence. In: Financial Management, 28, 75–88.

ADMATI, A. R. AND PFLEIDERER, P. (1988): A Theory of Intraday Patterns: Volume and Price Variability. In: Review of Financial Studies, 1, 3–40.

ADMATI, A. R. AND PFLEIDERER, P. (1989): Divide and Conquer: A Theory of Intraday and Day-of-the-Week Mean Effects. In: Review of Financial Studies, 2, 189–223.

AFFLECK-GRAVES, J., HEDGE, S. P. AND MILLER, R. E. (1994): Trading Mechanisms and the Components of the Bid-Ask Spread. In: Journal of Finance, 49, 1471–1488.

AKERLOF, G. A. (1970): The Market for "Lemons": Quality Uncertainty and the Market Mechanism. In: Quarterly Journal of Economics, 84, 488–500.

ALTMAN, E. I., (ed.) (1987): Handbook of Financial Markets and Institutions. Wiley Professional Banking and Finance Series. 6th edition, New York: John Wiley & Sons.

AMIHUD, Y. AND MENDELSON, H. (1980): Dealership Market: Market Making with Inventory. In: Journal of Financial Economics, 8, 31–53.

AMIHUD, Y. AND MENDELSON, H. (1986): Asset Pricing and the Bid-Ask Spread. In: Journal of Financial Economics, 17, 223–249.

ANGEL, J. J. (1997): How Best to Supply Liquidity to a Small-Capitalization Securities Market. Mimeo, Georgetown University.

ANGEL, J. J. AND WEAVER, D. G. (1998): Priority Rules !. Mimeo, Georgetown University.

ARROW, K. J. (1963): The Theory of Risk Aversion, chapter 9, 147–171. In: ARROW (1984).

ARROW, K. J. (1964): The Role of Securities in the Optimal Allocation of Risk-bearing. In: Review of Economic Studies, 31, 91–96.

ARROW, K. J. (1984): Collected Papers of Kenneth J. Arrow: Individual Choice under Certainty and Uncertainty. Oxford: Basil Blackwell.

ASCHINGER, G. (1990): General Remarks on Portfolio Theory and its Application. Working Paper Nr. 158, Insitute for Economic and Social Sciences, University of Fribourg (Switzerland).

AUMANN, R. J. AND SHAPLEY, L. S. (1976): Long-Term Competition - A Game Theoretic Analysis, 1–15. In: MEGIDDO (1974).

AXELROD, R. (1984): The Evolution of Cooperation. New York: Basic Books, Inc.

BACK, K., CAO, H. H. AND WILLARD, G. A. (2000): Imperfect Competition among Informed Traders. In: Journal of Finance, 55, 2117–2155.

BAGEHOT, W. (1971): The Only Game in Town. In: Financial Analysts Journal, 27, 12–14, 22.

BAMBERG, G. AND BAUR, F. (1993): Statistik. 8 edition, Munich: Oldenbourg Verlag.

BARCLAY, M. J. (1997): Bid-Ask Spreads and the Avoidance of Odd-Eighths Quotes on NASDAQ: An Examination of Exchange Listings. In: Journal of Financial Economics, 45, 35–60.

BARCLAY, M. J., CHRISTIE, W. G., HARRIS, J. H., KANDEL, E. AND SCHULTZ, P. H. (1999): Effects of Market Reform on the Trading Costs and Depths of NASDAQ Stocks. In: Journal of Finance, 59, 1–34.

BAUMOL, W. J., PANZAR, J. C. AND WILLIG, R. D. (1988): Contestable Markets and the Theory of Industry Structure. Revised edition, San Diego, CA: Harcourt Brace Jovanovich Publishers.

BEED, C. AND BEED, C. (2000): Intellectual Progress and Academic Economics: Rational Choice and Game Theory. In: Journal of Postkeynesian Economics, 23, 163–185.

BEEKMANN, W. (1995): Analysis I. Script for Course 01132 at the FernUniversität Hagen, Germany.

BEEKMANN, W. (1996): Analysis II. Script for Course 01133 at the FernUniversität Hagen, Germany.

BENHAMOU, E. AND SERVAL, T. (2000): On the Competition Between ECNs, Stock Markets and Market Makers. London School of Economics Financial Markets Groups Discussion Paper Series No. 345.

BENOIT, J.-P. AND KRISHNA, V. (1985): Finitely Repeated Games. In: Econometrica, 53, 890–904.

BENOIT, J.-P. AND KRISHNA, V. (1987): Nash Equilibria of Finitely Repeated Games. In: International Journal of Game Theory, 16, 197–204.

BERNHARDT, D. AND HUGHSON, E. (1996): Discrete Pricing and the Design of Dealership Markets. In: Journal of Economic Theory, 71, 148–182.

BERNHARDT, D. AND HUGHSON, E. (1997): Splitting Orders. In: Review of Financial Studies, 10, 69–101.

BERNHEIM, B. D. AND WHINSTON, M. D. (1990): Multimarket Contact and Collusive Behavior. In: RAND Journal of Economics, 21, 1–26.

BERTRAND, J. (1883): Théorie mathématique de la richesse sociale. In: Journal des Savants, 499–508.

BESSEMBINDER, H. (1999): Trade Execution Costs on NASDAQ and the NYSE: A Post-Reform Comparison. In: Journal of Financial and Quantitative Analysis, 34, 387–407.

BIAIS, B. (1993): Price Formation and Equilibrium Liquidity in Fragmented and Centralized Markets. In: Journal of Finance, 58, 157–185.

BIAIS, B., FOUCAULT, T. AND SALANI, F. (1995): Implicit Collusion on Wide Spreads. Working Paper No. 153, Finance and Banking Discussion Paper Series, Universitat Pompeu Fabra.

BLOOMFIELD, R. AND O'HARA, M. (1999): Market Transparency: Who Wins and Who Loses. In: Review of Financial Studies, 12, 5–35.

BLOOMFIELD, R. AND O'HARA, M. (2000): Can transparent markets survive?. In: Journal of Financial Economics, 55, 425–459.

BLUME, L., EASLEY, D. AND O'HARA, M. (1994): Market Statistics and Technical Analysis: The Role of Volume. In: Journal of Finance, 49, 153–181.

BONDARENKO, O. (1999): Competing Market Makers, Liquidity Provision, and Bid-Ask Spread. Mimeo, University of Illinois at Chicago.

BORDER, K. C. (1985): Fixed Point Theorems with Applications to Economics and Game Theory. Cambridge: Cambridge University Press.

BRACHINGER, H.-W. AND WEBER, M. (1997): Risk as a Primitive: A Survey of Measures of Perceived Risk. In: OR Spektrum, 19, 235–250.

BRENNAN, M. J. AND SUBRAHMANYAM, A. (1996): Market Microstructure and Asset Pricing: On the Compensation for Illiquidity in Stock Returns. In: Journal of Financial Economics, 41, 441–464.

BROOKS, R. AND MASSON, J. (1996): Performance of Stoll's Spread Component Estimator: Evidence from Simulations, Time Series, and Cross-Sectional Data. In: Journal of Financial Research, 19, 459–476.

BROUWER, L. E. J. (1912): Über Abbildung von Mannigfaltigkeiten. In: Mathematische Annalen, 71, 97–115.

BROWN, C. (2000): Electronic Communications Networks - Der Einfluss der Informationstechnologie auf die Organisation von Wertpapiermärkten. Master's thesis, Zurich, University of Zurich.

BROWN, D. P. AND JENNINGS, R. H. (1989): On Technical Analysis. In: Review of Financial Studies, 2, 527–551.

BROWN, D. P. AND ZHANG, Z. M. (1997): Market Orders and Market Efficiency. In: Journal of Finance, 52, 277–308.

BÜSCHGEN, H. (1991): Das kleine Börsen-Lexikon. 19th edition, Düsseldorf: Verlag Wirtschaft und Finanzen.

CABALLE, J. AND KRISHNAN, M. (1994): Imperfect Competition in a Multi-Security Market with Risk Neutrality. In: Econometrica, 62, 595–604.

CALCAGNO, R. AND LOVO, S. M. (1998): Bid-Ask Price Competition with Asymmetric Information between Market Makers. CORE Discussion Paper 9816, Universite de Catholique de Louvain.

CAMPBELL, J. Y., LO, A. W. AND MACKINLAY, A. C. (1997): The Econometrics of Financial Markets. Princeton, NJ: Princeton University Press.

CASON, T. N. (2000): The Opportunity for Conspirancy in Asset Markets Organized with Dealer Intermediaries. In: Review of Financial Studies, 13, 385–416.

CHAKRAVARTY, S. AND HOLDEN, C. W. (1995): An Integrated Model of Market and Limit Orders. In: Journal of Financial Intermediation, 4, 213–241.

CHAKRAVARTY, S., SARKAR, A. AND WU, L. (1997): Estmating the Adverse Selection Costs in Markets with Multiple Informed Traders. Federal Reserve Bank of New York Research Paper #9713.

CHAU, M. (1998): Dynamic Trading and Market-Making with Inventory Costs and Private Information. Mimeo, ESSEC, France.

CHEN, H.-C. AND RITTER, J. R. (2000): The Seven Percent Solution. In: Journal of Finance, 55, 1105–1131.

CHORDIA, T., ROLL, R. AND SUBRAHMANYAM, A. (2000): Commonality in Liquidity. In: Journal of Financial Economics, 56, 3–28.

CHRISTIE, W. G., HARRIS, J. AND SCHULTZ, P. H. (1994): Why Did NASDAQ Market Makers Stop Avoiding Odd-Eighth Quotes?. In: Journal of Finance, 49, 1841–1860.

CHRISTIE, W. G. AND SCHULTZ, P. H. (1994): Why Do NASDAQ Market Makers Avoid Odd-eighth Quotes?. In: Journal of Finance, 49, 1813–1840.

CHRISTIE, W. G. AND SCHULTZ, P. H. (1999): The Initiation and Withdrawal of Odd-eighth Quotes among NASDAQ Stocks: An Empirical Analysis. In: Journal of Financial Economics, 52, 409–442.

CHUNG, K. H., VAN NESS, B. F. AND VAN NESS, R. A. (1999a): Limit Orders and the Bid-Ask Spread. In: Journal of Financial Economics, 53, 255–287.

CHUNG, K. H., VAN NESS, B. F. AND VAN NESS, R. A. (1999b): Spreads, Depths, and Quote Clustering on the NYSE and NASDAQ: Evidence after the 1997 SEC Rule Changes. In: Journal of Business, under review.

COCHRANE, J. (2000): Asset Pricing. Chicago: Graduate School of Business, University of Chicago.

COHEN, K. J., HAWAWINI, G., MAIER, S. F., SCHWARTZ, R. A. AND WHITCOMB, D. K. (1980): Implications of Microstructure Theory for Empirical Research on Stock Price Behavior. In: Journal of Finance, 35, 249–257.

COHEN, K. J., MAIER, S. F., SCHWARTZ, R. A. AND WHITCOMB, D. K. (1981): Transactions Costs, Order Placement Strategy, and Existence of the Bid-Ask Spread. In: Journal of Political Economy, 89, 287–305.

COHEN, K. J., MAIER, S. F., SCHWARTZ, R. A. AND WHITCOMB, D. K. (1986): The Microstructure of Securities Markets. Englewood Cliffs, NJ: Prentice-Hall.

COONEY, J. W., VAN NESS, B. F. AND VAN NESS, R. A. (1999): Do Investors Avoid Odd-Eights Prices? Evidence from NYSE Limit Orders. Mimeo, University of Kentucky.

COPELAND, T. E. AND GALAI, D. (1983): Information Effects on the Bid-Ask-Spread. In: Journal of Finance, 38, 1457–1469.

CORDELLA, T. AND FOUCAULT, T. (1999): Minimum Price Variations, Time Priority, and Quote Dynamics. In: Journal of Financial Intermediation, 8, 141–173.

COUGHENOUR, J. AND SHASTRI, K. (1999): Symposium on Market Microstructure: A Review of Empirical Research. In: Financial Review, 34(4), 1–28.

COURNOT, A. (1963 (orig. 1838)): The Mathematical Principles of the Theory of Wealth. Homewood, Ill.: Richard D. Irwin, Inc.

CRAWFORD, V. P. AND HALLER, H. (1990): Learning to Cooperate: Optimal Play in Repeated Coordination Games. In: Econometrica, 58, 571–595.

CYMBALISTA, F. (1998): Zur Unmöglichkeit rationaler Bewertung unter Unsicherheit - Eine monetär-keynesianische Kritik der Diskussion um die Markteffizienzthese. Studien zur Monetären Ökonomie. Marburg: Metropolis Verlag.

DASGUPTA, P. AND MASKIN, E. (1986): The Existence of Equilibrium in Discontinuous Economic Games. I: Theory. In: Review of Economic Studies, 53, 1–26.

DE LA VEGA, J. (1994 (orig. 1688)): Confusion de Confusiones (Die Verwirrung der Verwirrungen). Kulmbach, Germany: Boersenbuch-Verlag.

DEMSETZ, H. (1968): The Cost of Transacting. In: Quarterly Journal of Economics, 82, 33–53.

DEMSETZ, H. (1997): Limit Orders and the Alleged NASDAQ Collusion. In: Journal of Financial Economics, 45, 91–95.

DENNERT, J. (1993): Price Competition between Market Makers. In: Review of Economic Studies, 60, 735–751.

DEUTSCHE BÖRSE GROUP, (ed.) (1999): Rules and Regulations. Frankfurt.

DIAMOND, D. W. AND VERRECCHIA, R. E. (1981): Information Aggregation in a Noisy Rational Expectations Economy. In: Journal of Financial Economics, 9, 221–235.

DOMOWITZ, I. (1993): A Taxonomy of Automated Trade Execution Systems. In: Journal of International Money and Finance, 12, 607–631.

DORAN, L. (1999): Market Making in the Third Market for NYSE-Listed Securities. In: Financial Review, 34(4), 29–54.

DORAN, L., LEHN, K. AND SHASTRI, K. (1995): Do NASDAQ Market Makers Collude? Evidence from 19c-3 Stocks. Mimeo, University of Pittsburgh, Katz School of Business.

DUFFIE, D. (1996): Dynamic Asset Pricing. 2nd edition, Princeton, NJ: Princeton University Press.

DUMAS, B. AND ALLAZ, B. (1996): Financial Securities: Market Equilibrium and Pricing Methods. London: Chapman & Hall.

DUTTA, P. K. (1995a): A Folk Theorem for Stochastic Games. In: Journal of Economic Theory, 66, 1–32.

DUTTA, P. K. (1995b): Collusion, Discounting and Dynamic Games. In: Journal of Economic Theory, 66, 289–306.

DUTTA, P. K. AND MADHAVAN, A. (1997): Competition and Collusion in Dealer Markets. In: Journal of Finance, 52, 245–276.

EASLEY, D. AND O'HARA, M. (1991): Order Form and Information in Securities Markets. In: Journal of Finance, 46, 905–927.

EASLEY, D. AND O'HARA, M. (1995): Market Microstructure, chapter 12, 357–383. In: JARROW ET AL. (1995).

ECONOMIDES, N. AND SCHWARTZ, R. A. (1995): Equity Trading Practices and Market Structure: Assessing Asset Managers' Demand for Immediacy. In: Financial Markets, Institutions & Instruments, 4(4), 1–46.

EDGEWORTH, F. Y. (1897): La Teoria pura del monopolio. In: Giornale degli Economisti, 13–31.

FAMA, E. F. (1970): Efficient Capital Markets: A Review of Theory and Empirical Work. In: Journal of Finance, 25, 383–423.

FLOOD, M. D., HUISMAN, R., KOEDIJK, K. G. AND LYONS, R. K. (1998): Search Costs: The Neglected Spread Component. Mimeo, Haas School of Business, University of California at Berkeley.

FLOOD, M. D., HUISMAN, R., KOEDIJK, K. G. AND MAHIEU, R. J. (1999): Quote Disclosure and Price Discovery in Multiple-Dealer Financial Markets. In: Review of Financial Studies, 12, 37–59.

FOSTER, F. D. AND VISWANATHAN, S. (1990): A Theory of Interday Variations in Volume, Variance, and Trading Costs in Securities Markets. In: Review of Financial Studies, 3, 593–624.

FOSTER, F. D. AND VISWANATHAN, S. (1993a): The Effect of Public Information and Competition on Trading Volume and Price Volatility. In: Review of Financial Studies, 6, 23–56.

FOSTER, F. D. AND VISWANATHAN, S. (1993b): Variations in Trading Volume, Return Volatility, and Trading Costs: Evidence on Recent Price Formation Models. In: Journal of Finance, 48, 187–211.

FOUCAULT, T. AND PARLOUR, C. A. (1999): Competition for Listings. CEPR Discussion Series No. 2222.

FREY, B. S. (2000): Was bewirkt die Volkswirtschaftslehre?. In: Perspektiven der Wirtschaftspolitik, 1(1), 5–33.

FRIEDMAN, B. M. (1979): Optimal Expectations and the Extreme Information Assumptions of Rational Expectations Macromodels. In: Journal of Monetary Economics, 5, 23–41.

FRIEDMAN, J. (1971): A Noncooperative Equilibrium for Supergames. In: Review of Economic Studies, 38, 1–12.

FUDENBERG, D. AND LEVINE, D. (1983): Subgame-perfect Equilibria of Finite and Infinite Horizon Games. In: Journal of Economic Theory, 31, 227–256.

FUDENBERG, D. AND MASKIN, E. (1986): The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. In: Econometrica, 54, 533–556.

FUDENBERG, D. AND MASKIN, E. (1990): Nash and Perfect Equilibria of Discounted Repeated Games. In: Journal of Economic Theory, 50, 194–206.

FUDENBERG, D. AND TIROLE, J. (1991): Game Theory. Cambridge, Mass.: MIT Press.

FURBUSH, D. AND SMITH, J. W. (1996): Quoting Behavior on NASDAQ: The Effects of Clustering on Relative Spreads. NASD Working Paper 96-01.

GARBADE, K. D. AND SILBER, W. L. (1979): Structural Organization of Secondary Markets: Clearing Frequency, Dealer Activity and Liquidity Risk. In: Journal of Finance, 34, 577–593.

GARFINKEL, J. A. AND NIMALENDRAN, M. (1998): Market Structure and Trader Anonymity: An Analysis of Insider Trading. Mimeo, University of Florida.

GARMAN, M. B. (1976): Market Microstructure. In: Journal of Financial Economics, 3, 257–275.

GARVEY, G. T. AND MCCORRY, M. S. (1997): Multidimensional Competition on NASDAQ: Did Traders Gain When Dealers Stopped Avoiding Odd-Eighth Quotes?. UBC-FIN 97-3, University of British Columbia, Vancouver, Canada.

GEHRIG, T. (1998): Competing Markets. In: European Economic Review, 42, 277–310.

GEHRIG, T. AND JACKSON, M. (1998): Bid-Ask Spreads with Indirect Competition Among Specialists. In: Journal of Financial Markets, 1, 89–119.

GEORGE, T. J., KAUL, G. AND NIMALENDRAN, M. (1991): Estimation of the Bid-Ask Spread and its Components: A New Approach. In: Review of Financial Studies, 4, 623–656.

GLICKSBERG, I. L. (1952): A further Generalization of the Kakutani Fixed Point Theorem with Application to Nash Equilibrium Points. In: Proceedings of the National Academy of Sciences, 38, 170–174.

GLOSTEN, L. R. (1989): Insider Trading, Liquidity, and the Role of the Monopolist Specialist. In: Journal of Business, 62, 211–235.

GLOSTEN, L. R. (1994): Is the Electronic Limit Order Book Inevitable?. In: Journal of Finance, 49, 1127–1161.

GLOSTEN, L. R. AND HARRIS, L. E. (1988): Estimating the Components of the Bid/Ask Spread. In: Journal of Financial Economics, 21, 123–142.

GLOSTEN, L. R. AND MILGROM, P. R. (1985): Bid, Ask and Transaction Prices in a Specialist Market with Heterogenously Informed Traders. In: Journal of Financial Economics, 14, 71–100.

GODEK, P. E. (1996): Why NASDAQ Market Makers Avoid Odd-Eighths Quotes. In: Journal of Financial Economics, 41, 465–474.

GOLDSTEIN, M. A. AND KAVAJECZ, K. A. (2000): Eighths, Sixteenths, and Market Depth: Changes in Tick Size and Liquidity Provision on the NYSE. In: Journal of Financial Economics, 56, 125–149.

GOMBER, P. (2000): Elektronische Handelssysteme. Innovative Konzepte und Technologien im Wertpapierhandel. Heidelberg: Physica-Verlag.

GOMES, G. M. (1982): Irrationality of Rational Expectations. In: Journal of Postkeynesian Economics, 5, 51–65.

GREEN, E. J. AND PORTER, R. H. (1984): Noncooperative Collusion under Imperfect Price Information. In: Econometrica, 52, 87–100.

GREENE, W. H. (1997): Econometric Analysis. Upper Saddle River, NJ: Prentice-Hall, Inc.

GROSSMAN, S. J. AND MILLER, M. H. (1988): Liquidity and Market Structure. In: Journal of Finance, 43, 617–637.

GROSSMAN, S. J. AND STIGLITZ, J. E. (1980): On the Impossibility of Informationally Efficient Markets. In: American Economic Review, 70, 393–408.

HAGERTY, K. (1991): Equilibrium Bid-Ask Spreads in Markets with Multiple Assets. In: Review of Economic Studies, 58, 237–257.

HANSCH, O., NAIK, N. Y. AND VISWANATHAN, S. (1998): Do Inventories Matter in Dealership Markets? Evidence from the London Stock Exchange. In: Journal of Finance, 53, 1623–1656.

HANSCH, O., NAIK, N. Y. AND VISWANATHAN, S. (1999): Preferencing, Internalization, Best Execution, and Dealer Profits. In: Journal of Finance, 54, 1799–1828.

HANSEN, R. S. (2000): Do Investment Banks Compete in IPOs?: The Advent of the '7% plus Contract'. In: Journal of Financial Economics, 59, forthcoming.

HARRIS, L. (1991): Stock Price Clustering and Discreteness. In: Review of Financial Studies, 4, 389–415.

HARRIS, L. E. (1994): Minimum Price Variations, Discrete Bid-Ask Spreads, and Quotation Sizes. In: Review of Financial Studies, 7, 149–178.

HARRIS, L. E. AND HASBROUCK, J. (1996): Market vs. Limit Orders: The SuperDOT Evidence on Order Submission Strategy. In: Journal of Financial and Quantitative Analysis, 31, 213–231.

HARSANYI, J. C. (1967-1968): Games with Incomplete Information Played by Bayesian Players. In: Management Science, 14, 159–182, 320–344, 486–502.

HASBROUCK, J. (1988): Trades, Quotes, Inventories, and Information. In: Journal of Financial Economics, 22, 229–252.

HAYEK, F. A. (1945): The Use of Knowledge in Society. In: American Economic Review, 35, 519–530.

HE, Y. AND WU, C. (1999): The Effects of Market Reform on the Informed Trading Costs of NASDAQ Stocks. Mimeo, San Francisco State University.

HEIDLE, H. G. AND HUANG, R. D. (1999): Information-Based Trading in Dealer and Auction Markets: An Analysis of Exchange Listings. Mimeo, Owen Graduate School of Management, Vanderbilt University.

HENDERSHOTT, T. AND MENDELSON, H. (2000): Crossing Networks and Dealer Markets: Competition and Performance. In: Journal of Finance, 55, 2071–2115.

HO, T. AND STOLL, H. R. (1980): On Dealer Markets under Competition. In: Journal of Finance, 35, 259–268.

HO, T. AND STOLL, H. R. (1981): Optimal Dealer Pricing under Transactions and Return Uncertainty. In: Journal of Financial Economics, 9, 47–73.

HO, T. AND STOLL, H. R. (1983): The Dynamics of Dealer Markets under Competition. In: Journal of Finance, 38, 1053–1074.

HOLDEN, C. W. AND SUBRAHMANYAM, A. (1992): Long-Lived Private Information and Imperfect Competition. In: Journal of Finance, 47, 247–270.

HOPT, K. J. AND BAUM, H. (1997): Börsenrechtsreform in Deutschland, chapter 3, 287–467. In: HOPT ET AL. (1997).

HOPT, K. J., RUDOLPH, B. AND BAUM, H., (eds.) (1997): Börsenreform - Eine ökonomische, rechtsvergleichende und rechtspolitische Untersuchung. Stuttgart: Schäffer-Poeschel Verlag.

HUANG, R. D. (2000): Price Discovery by ECNs and NASDAQ Market Makers. Mimeo, Owen Graduate School of Business, Vanderbilt University.

HUANG, R. D. AND STOLL, H. R. (1997): The Components of the Bid-Ask Spread: A General Approach. In: Review of Financial Studies, 10, 995–1034.

HUANG, R. D. AND STOLL, H. R. (1999): Tick Size, Bid-Ask Spreads and Market Structure. Owen Graduate School of Business, Vanderbilt University, Working Paper 99-05.

HUCK, S., NORMANN, H.-T. AND OECHSSLER, J. (2000): Does Information about Competitors' Actions Increase or Decrease Competition in Experimental Oligopoly Markets?. In: International Journal of Industrial Organization, 18, 39–57.

INGERSOLL, J. E. (1987): Theory of Financial Decision Making. Studies in Financial Economics. Savage, MD: Rowman & Littlefield.

INTRILIGATOR, M. D. (1971): Mathematical Optimization and Economic Theory, chapter 13, 326–343. Englewwod Cliffs, N. J., USA: Prentice Hall.

ITÔ, K. (1951): On Stochastic Differential Equations. In: Memoirs of the American Mathematical Society, 4, 1–51.

JAIN, N. AND MIRMAN, L. J. (1999): Insider Trading with Correlated Signals. In: Economics Letters, 65, 105–113.

JANSSEN, M. C. W. (1998): Focal Points, 150–155. Vol. 2, in: NEWMAN (1998).

JARROW, R. A., MAKSIMOVIC, V. AND ZIEMBA, W. T., (eds.) (1995): Finance. Handbooks in Operations Research and Management Science. Amsterdam: Elsevier.

JEVONS, W. S. (1924 (orig. 1911)): Die Theorie der Politischen Ökonomie. Jena: Verlag von Gustav Fischer.

JONES, C. M. AND LIPSON, M. L. (2000): Sixteenths: Direct Evidence on Institutional Execution Costs. In: Journal of Financial Economics, 59, fortcoming.

KAKUTANI, S. (1941): A Generalization of Brouwer's Fixed Point Theorem. In: Duke Mathematical Journal, 8, 416–427.

KANDEL, E. AND MARX, L. (1997): NASDAQ Market Structure and Spread patterns. In: Journal of Financial Economics, 45, 61–89.

KANDEL, E. AND MARX, L. (1999a): Payments for Order Flow on NASDAQ. In: Journal of Finance, 59, 35–66.

KANDEL, E. AND MARX, L. M. (1999b): Odd-Eighth Avoidance as a Defense against SOES bandits. In: Journal of Financial Economics, 51, 85–102.

KANDORI, M. AND MATSUSHIMA, H. (1998): Private Observation, Communication and Collusion. In: Econometrica, 66, 627–652.

KEENAN, W. M. (1987): The Securities Industry: Securities Trading and Investment Banking, chapter 9. In: ALTMAN (1987).

KEYNES, J. M. (1930): A Treatise on Money. London.

KEYNES, J. M. (1936): The General Theory of Employment, Interest, and Money. London.

KLEIDON, A. W. AND WILLIG, R. (1995): Why Do Christie and Schultz Infer Collusion from Their Data?. Mimeo, Princeton University.

KLOCK, M. AND MCCORMICK, D. T. (1999): The Impact of Market Maker Competition on NASDAQ Spreads. In: Financial Review, 34(4), 55–74.

KNIGHT, F. H. (1957 (Orig. 1921)): Risk, Uncertainty and Profit. New York: Kelly & Millman, Inc.

KRAHNEN, J. P. AND WEBER, M. (1997): Market Making in the Laboratory: Does Competition Matter?. Johann-Wolfgang von Goethe- Universität Frankfurt, Fachbereich Wirtschaftswissenschaften, Working Paper Series: Finance and & Accounting, No. 4.

KRAUSE, A. (1999): Microstructure Effects on Daily Return Volatility in Financial Markets. Mimeo, University of Fribourg.

KREPS, D. M. (1998): Bounded Rationality, 168–173. Vol. 1, in: NEWMAN (1998).

KREPS, D. M., MILGROM, P., ROBERTS, J. AND WILSON, R. (1982): Rational Cooperation in the Finitely Repeated Prisoners' Dilemma. In: Journal of Economic Theory, 27, 245–252.

KREPS, D. M. AND WILSON, R. (1982): Reputation and Imperfect Information. In: Journal of Economic Theory, 27, 253–279.

KRISHNAN, M. (1992): An Equivalence between the Kyle (1985) and the Glosten-Milgrom (1985) Models. In: Economics Letters, 40, 333–338.

KYLE, A. S. (1985): Continuous Auctions and Insider Trading. In: Econometrica, 53, 1315–1335.

KYLE, A. S. (1989): Informed Speculation with Imperfect Competition. In: Review of Economic Studies, 56, 317–356.

LAUX, P. A. (1993): Trade Sizes and Theories of the Bid-Ask Spread. In: Journal of Financial Research, 16, 237–249.

LAUX, P. A. (1995): Dealer Market Structure, Outside Competition, and the Bid-Ask Spread. In: Journal of Economic Dynamics and Control, 19, 683–710.

LEE, R. (1998): What is an Exchange? The Automation, Management and Regulation of Financial Markets. Oxford: Oxford University Press.

LEHRER, E. AND PAUZNER, A. (1999): Repeated Games with Differential Time Preferences. In: Econometrica, 67, 393–412.

LESMOND, D. A., OGDEN, J. P. AND TRZCINKA, C. A. (1999): A New Estimate of Transaction Costs. In: Review of Financial Studies, 12, 1113–1141.

LEVY, H. AND SARNAT, M. (1972): Investment and Portfolio Analysis. New York: John Wiley & Sons, Inc.

LIN, J.-C., SANGER, G. C. AND BOOTH, G. G. (1995): Trade Size and Components of the Bid-Ask Spread. In: Review of Financial Studies, 8, 1153–1183.

LINTNER, J. (1965): Security Prices, Risk, and Maximal Gains from Diversification. In: Journal of Finance, 20, 587–615.

LO, A. W., MACKINLAY, A. C. AND ZHANG, J. (2000): Econometric Models of Limit-Order Executions. In: Journal of Financial Economics, forthcoming.

LUCAS, R. E. AND PRESCOTT, E. C. (1971): Investment under Uncertainty. In: Econometrica, 39, 659–681.

MACDONALD, J. M. (2000): Demand, Information, and Competition: Why Do Food Prices Fall at Seasonal Demand Peaks?. In: Journal of Industrial Economics, 48, 27–45.

MADHAVAN, A. (1992): Trading Mechanisms in Securities Markets. In: Journal of Finance, 47, 607–641.

MADHAVAN, A., RICHARDSON, M. AND ROOMANS, M. (1997): Why Do Securities Prices Change? A Transaction Level Analysis of NYSE Stocks. In: Review of Financial Studies, 10, 1035–1064.

MARKOWITZ, H. M. (1959): Portfolio Selection: Efficient Diversification of Investments. New Haven, CT: Yale University Press.

MARQUARDT, D. S. (1998): Financial Markets Performance: Theory and Empirical Evidence. Doctoral Dissertation Nr. 2142, St. Gallen, University of St. Gallen.

MARSDEN, J. AND WEINSTEIN, A. (1985): Calculus. 2nd edition, New York: Springer Verlag.

MARSHALL, A. (1997, orig. 8th. ed. 1920): Principles of Economics. Amherst, N.Y.: Prometheus Books.

MATSUSHIMA, H. (1998): Multimarket Contact, Imperfect Monitoring, and Implicit Collusion. Discussion Paper CIRJE-F-24, Faculty of Economics, University of Tokyo.

MEGIDDO, N., (ed.) (1974): Essays in Game Theory in Honor of Michael Maschler. Hamburg: Springer Verlag.

MERKT, H. (1997): Zur Entwicklung des deutschen Börsenrechts von den Anfängen bis zum Zweiten Finanzmarktförderungsgesetz, 17–141. In: HOPT ET AL. (1997).

MILGROM, P. AND ROBERTS, J. (1982): Predation, Reputation and Entry Deterrence. In: Journal of Economic Theory, 27, 280–312.

MOULIN, H. (2000): Priority Rules and Other Asymmetric Rationing Methods. In: Econometrica, 68, 643–684.

MUTH, J. F. (1961): Rational Expectations and the Theory of Price Movements. In: Econometrica, 29, 315–335.

NAIK, N. Y. AND YADAV, P. K. (1999): The Effects of Market Reform on Trading Costs of Public Investors: Evidence from the London Stock Exchange. Mimeo, London Business School.

NASD, INC. (1991): Inducements for Order Flow. Report to the NASD Board of Governors.

NASD, INC. (1997): Market Quality Monitoring.

NASD, INC. (1998): Manual. Chicago: CCH Inc.

NASD, INC. (2000): Highlights of the Recent Economic Advisory Board Meeting. In: Research Matters, 3(1), 9.

NASH, J. (1950): Equilibrium Points in N-person Games. In: Proceedings of the National Academy of Sciences, 36, 48–49.

NASH, J. (1951): Non-Cooperative Games. In: Annals of Mathematics, 54, 286–295.

NEAL, R. AND WHEATLEY, S. (1995): How Reliable are Adverse Selection Models of the Bid-Ask Spread?. Federal Reserve Bank of Kansas City Research Working Paper RWP 95-02.

NELLING, E. F. AND GOLDSTEIN, M. A. (1999): Market Making and Trading in NASDAQ Stocks. In: Financial Review, 34(1), 27–44.

NEWMAN, P., (ed.) (1998): The New Palgrave Dictionary of Economics and the Law. London: Macmillan Reference Ltd.

NEYMAN, A. (1999): Cooperation in Repeated Games when the Number of Stages is not Commonly Known. In: Econometrica, 67, 45–64.

NILSSON, A. (1999): Transparency and Competition. Mimeo, Stockholm School of Economics.

O'HARA, M. (1995): Market Microstructure Theory. Cambridge, Mass.: Blackwell Publishers.

PAGANO, M. AND ROËLL, A. (1996): Transparency and Liquidity: A Comparison of Auction and Dealer Markets with Informed Trading. In: Journal of Finance, 51, 579–611.

PARLOUR, C. A. (1998): Price Dynamics in Limit Order Markets. In: Review of Financial Studies, 11, 789–816.

PESARAN, M. H. (1987): The Limits to Rational Expectations. Oxford: Basil Blackwell Ltd.

PHLIPS, L. (1995): Competition Policy: A Game Theoretic Perspective. Cambridge: Cambridge University Press.

PIRRONG, C. (1999): The Organization of Financial Exchange Markets: Theory and Evidence. In: Journal of Financial Markets, 2, 329–357.

PORTER, D. C. AND WEAVER, D. G. (1996): Estimating Bid-Ask Spread Components: Specialist versus Multiple Market Maker Systems. In: Review of Quantitative Finance and Accounting, 6, 167–180.

PRATT, J. W. (1964): Risk Aversion in the Small and in the Large. In: Econometrica, 32, 122–136.

RAMANLAL, P. (1999): Liquidity Trading in Market Microstructure Theory. In: Review of Quantitative Finance and Accounting, 13, 29–38.

REDMAN, D. A. (1992): A Reader's Guide to Rational Expectations. Aldershot, Great Britain: Edward Elgar Publishing Ltd.

RICKER, J. P. (1997): Decimals for NASDAQ. Available at http://www.electronic-traders.org/acad006.html.

RICKER, J. P. (1998): The NYSE vs. NASDAQ Market Structure. Available at http://www.electronic-traders.org/acad007.html.

RICKER, J. P. (1999): Breaking the Eighth: Sixteenths on the New York Stock Exchange. Available at http://www.electronic-traders.org/acad008.html.

ROLL, R. (1984): A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market. In: Journal of Finance, 39, 1127–1139.

RUDOLPH, B. AND RÖHRL, H. (1997): Grundfragen der Börsenorganisation aus ökonomischer Sicht, chapter 2, 143–285. In: HOPT ET AL. (1997).

SCHELLING, T. C. (1960): The Strategy of Conflict. Cambridge, Mass.: Harvard University Press.

SCHILLER, B. AND MAREK, M. (2000): DTB + SOFFEX = EUREX. In: WiSt, (4), 219–221.

SCHMIDT, H. AND TRESKE, K. (1996): Komponenten der Geld-Brief-Spanne am Deutschen Aktienmarkt. In: Zeitschrift für Betriebswirtschaft, 66, 1033–1056.

SCHULTZ, P. H. (2000): Regulatory and Legal Pressures and the Costs of NASDAQ Trading. In: Review of Financial Studies, 13, 917–957.

SCHUMANN, J. (1992): Grundzüge der mikroökonomischen Theorie. 6th edition, Berlin: Springer Verlag.

SCHWARTZ, R. A. (1988): Equity Markets: Structure, Trading, and Performance. New York: Harper & Row.

SCHWARTZ, R. A. (1993): Reshaping the Equity Markets: A Guide For the 1990s. Homewood, Ill.: Business One Irwin.

SECURITIES AND EXCHANGE COMMISSION (2000a): Order Directing the Exchanges and NASD to Submit a Phase-In Plan to Implement Decimal Pricing in Equity Securities and Option, Release No. 34-42914.

SECURITIES AND EXCHANGE COMMISSION (2000c): Release No. 34-42450; File No. SR-NYSE-99-48.

SECURITIES AND EXCHANGE COMMISSION (2000b): SEC Requests Comments on NYSE Rescission of Rule 390 and on Market Fragmentation. Press Release 2000-14, February 23, 2000.

SELTEN, R. (1965): Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit. In: Zeitschrift für die gesamte Staatswissenschaft, 12, 301–324.

SHARPE, W. F. (1970): Portfolio Theory and Capital Markets. New York: McGraw-Hill Book Co.

SHEFFRIN, S. M. (1996): Rational Expectations. Cambridge: Cambridge University Press.

SHLEIFER, A. (2000): Inefficient Markets - An Introduction to Behavorial Finance. Oxford: Oxford University Press.

SIMAAN, Y., WEAVER, D. G. AND WHITCOMB, D. K. (2000): The Quotation Behavior of ECNs and NASDAQ Market Makers. In: Journal of Financial Economics, under review.

SMITH, J. W., SELWAY, J. P. AND MCCORMICK, D. T. (1998): The Nasdaq Stock Market: Historical Background and Current Operation. NASD Working Paper 98-01.

SNELL, A. AND TONKS, I. (1998): Testing for Asymmetric Information and Inventory Control Effects in Market Maker Behaviour on the London Stock Exchange. In: Journal of Empirical Finance, 5, 1–25.

SNELL, A. AND TONKS, I. (1999): Measuring Microstructure Effects for Less-Liquid Stocks in a Dealer Market. University of Bristol, Mimeo.

SPAGNOLO, G. (1999): On Interdependent Supergames: Multimarket Contact, Concavity, and Collusion. In: Journal of Economic Theory, 89, 127–139.

SPIEGEL, M. AND SUBRAHMANYAM, A. (1992): Informed Speculation and Hedging in a Noncompetitive Securities Market. In: Review of Financial Studies, 5, 307–329.

SPULBER, D. F. (1999): Market Microstructure: Intermediaries and the Theory of the Firm. Cambridge: Cambridge University Press.

STOLL, H. R. (1978): The Supply of Dealer Services in Securities Markets. In: Journal of Finance, 33, 1133–1151.

STOLL, H. R. (1989): Inferring the Components of the Bid-Ask spread: Theory and Empirical Tests. In: Journal of Finance, 44, 115–134.

STOLL, H. R. (2000): Friction. In: Journal of Finance, 55, 1479–1514.

SUBRAHMANYAM, A. (1991): Risk Aversion, Market Liquidity, and Price Efficiency. In: Review of Financial Studies, 4, 417–441.

SUGDEN, R. (1995): A Theory of Focal Points. In: Economic Journal, 104, 533–550.

THALER, R. H., (ed.) (1993): Advances in Behavioral Finance. New York: Russel Sage Foundation.

THE NASDAQ STOCK MARKET, INC., (ed.) (1997): The NASDAQ Stock Market 1997 Factbook. New York.

THEISSEN, E. (1999): Floor versus Screen Trading: Evidence from the German Stock Market. Mimeo, Johann Wolfgang Goethe-Universität, Frankfurt.

THEPOT, J. AND THIETARD, R.-A., (eds.) (1991): Microeconomic Contributions to Strategic Management. Advanced Series in Management, Vol. 16. Amsterdam: North-Holland.

TIROLE, J. (1988): The Theory of Industrial Organization. Cambridge, MA: MIT Press.

TOBIN, J. (1958): Liquidity Preference as a Behavior Towards Risk. In: Review of Economic Studies, 25, 65–86.

TOBIN, J. (1966): The Theory of Portfolio Selection. In: F. H. Hahn and F. P. R. Brechling: The Theory of Interest Rates, New York, 3–51.

TONKS, I. (1997): The Equivalence of Screen Based Continuous-Auction and Dealer Markets. Mimeo, University of Bristol.

UNSER, M. AND OEHLER, A. (1998): Das Börsenhandelssystem XETRA. In: WiSt, (9), 463–468.

VAN NESS, B. F., VAN NESS, R. A. AND HSIEH, W.-L. (1999a): NASDAQ and The Chicago Stock Exchange: An Analysis of Multiple Market Trading. In: Financial Review, 34(4), 145–158.

VAN NESS, B. F., VAN NESS, R. A. AND PRUITT, S. W. (1999b): An Empirical Examination of the NASDAQ/CHX Dual-Trading Experiment. In: Financial Review, 34(3), 65–78.

VAN WEGBERG, M. AND VAN WITTELOOSTUIJN, A. (1991): Multimarket Competition: Entry Strategies and Entry Deterrence When the Entrant Has a Home Market, chapter 6, 93–119. In: THEPOT AND THIETARD (1991).

VAN WEGBERG, M. AND VAN WITTELOOSTUIJN, A. (1992a): Credible Entry Threats into Contestable Markets: A Symmetric Multi-Market Model of Contestability. In: Economica, 59, 437–452.

VAN WEGBERG, M. AND VAN WITTELOOSTUIJN, A. (1992b): Multimarket Competition: Theory and Evidence. In: Journal of Economic Behavior and Organization, 18, 273–282.

VISWANATHAN, S. AND WANG, J. J. D. (1999): Market Architecture: Limit Order Books versus Dealership Markets. Mimeo, Duke University.

VON NEUMANN, J. AND MORGENSTERN, O. (1953): The Theory of Games and Economic Behavior. Princeton: Princeton University Press.

VRIEND, N. J. (1996): A Model of Market Making. Economics Working Paper 184, Universitat Pompeu Fabra.

WAHAL, S. (1997): Entry, Exit, Market Makers, and the Bid-Ask Spread. In: Review of Financial Studies, 10, 871–901.

WESTON, J. P. (2000): Competition on the NASDAQ and the Impact of Recent Market Reforms. In: Journal of Finance, 55, fortcoming December.

WHITCOMB, D. K. (1997): Testimony before The House Committee on Commerce, Subcommittee on Finance. In: re H.R. 1053 "The Common Cents Stock Pricing Act of 1997".

WILSON, R. (1971): Computing Equilibria of N-Person Games. In: SIAM Journal of Applied Mathematics, 21, 80–87.

WU, W. AND JIANG, J.-H. (1962): Essential Equilibrium Points of N-Person Non-cooperative Games. In: Scientia Sinica, 11, 1307–1322.

YAVAS, A. (1992): Market Makers versus Match Makers. In: Journal of Financial Intermediation, 2, 33–58.