Limit Order Markets, Liquidity, and Price Impact

by

Ioanid Rosu
Ph.D. in Mathematics
Massachusetts Institute of Technology, 1999

B.A. and Diploma in Mathematics
University of Bucharest, 1994

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

June 2004

Signature of Author: .....
Sloan School of Management
April 26, 2004

Certified by .....
Andrew W. Lo
Harris and Harris Group Professor of Finance
Thesis Supervisor

Accepted by .....
Birger Wernerfelt
Director, Ph.D. Program, Sloan School of Management

Limit Order Markets, Liquidity, and Price Impact

by

Ioanid Rosu

Submitted to the Sloan School of Management
on April 26, 2004 in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

ABSTRACT

In this thesis, I explore various aspects of market liquidity and analyze its effect on asset prices. First, in a model of a limit order market I explain how to define liquidity and derive a price impact function. Second, I show how agents who have price impact generate a liquidity component in asset prices.

In Part I, I propose a continuous-time model of price formation in a limit-order market. Strategic liquidity traders arrive randomly to the market and dynamically choose between limit and market orders, trading off execution price with waiting costs. I prove the existence of a Markov equilibrium in which the bid and ask prices depend only on the numbers of buy and sell orders in the book, and which can be characterized in closed-form in several cases of interest. My model generates empirically verified implications for the shape of the limit-order book and the dynamics of prices and trades. In particular, I show that buy and sell orders can in some cases cluster away from the bid-ask spread, thus generating a concave price impact function.

In Part II, I lay the foundations for the model in Part I by explaining how to define multi-stage games with perfect information in continuous time. In this version, strategies are locally constant and have a finite number of jumps. Also, I allow for the possibility of "stopping the clock."

In Part III (joint with Andrew W. Lo and Jiang Wang), we analyze the effects of price impact in a rational infinite-horizon, discrete-time consumption–investment model. We assume that some agents' transactions change prices via an exogenously determined price impact function. Then the resulting equilibrium price, besides the risk premium, displays another component, which depends on the price impact coefficient. We interpret this component as a "liquidity premium."

Thesis Supervisor: Andrew W. Lo
Title: Harris and Harris Group Professor of Finance

ACKNOWLEDGEMENTS

There are many people who contributed, directly or indirectly, to the present thesis. Among them, I want to thank my advisor, Andrew Lo, first of all for encouraging and guiding my transition from mathematics to finance, and for his generous help and ideas throughout the Ph.D. program. I am also grateful to Jiang Wang for constant guidance and valuable insights, to Dimitri Vayanos and Xavier Gabaix for their encouragement and careful supervision.

I benefitted a lot from the intense research atmosphere at MIT and at the Sloan School in particular. And, of course, I had the privilege of receiving the support of many friends (too many to mention here) from the larger MIT community, without which my life here would not have been so rich and enjoyable.

Finally, this thesis is dedicated to the memory of my father. He was not only a loving parent, but also my mentor and best friend. His intellectual curiosity and high standards of research have been to me a constant source of inspiration and strength.

Contents

1 Introduction

Liquidity is an important concept in financial markets, and is frequently used by academics and practitioners alike. However, despite efforts of a large literature in market microstructure[1], liquidity has remained an ellusive concept. It is said that, like vulgarity, liquidity is easy to recognize, but difficult to define. At the same time, a growing literature in asset pricing[2] indicates that liquidity (defined typically as the price impact of a transaction) has a considerable effect on asset prices.

In this thesis, I explore both the nature of liquidity, and the means by which it affects asset prices. In the first line of inquiry, I consider a model of a specific type of market, called limit order market, where there are no market makers and trading is done according to a public order book. This simplifies the discussion, and allows me to give a definition of liquidity based on the arrival rates of the various types of traders, and their relative waiting costs. Using these basic notions, I derive an equilibrium order book and explain how it evolves in time. Then I define (in the context of limit order markets) both an instantaneous price impact function, as well as a price impact function relative to a given time interval.

This approach represents a departure from the typical market microstructure literature.[3] In that literature, liquidity is defined as the amount of adverse selection in the market. Prices change because suppliers of liquidity have to protect themselves from traders with superior information. In my model, prices change because the arrival of new agents changes the balance between various suppliers of liquidity.

In the second line of inquiry, I assume as given the price impact function, together with the corresponding time interval. Then in a discrete-time consumption–investement model, I show that if the agent who wants to trade has price impact, then asset prices should display, besides a risk component, a component which depends on the price impact function. This can be interpreted as a "liquidity premium," and provides a theoretical justification for the empirical findings mentioned above, which show that liquidity can significantly impact equilibrium prices.

[1] See the survey book by O'Hara (1995), ch. 8.

[2] Brennan and Subrahmanyam (1996), Pastor and Stambaugh (2003), etc.

[3] However, in recent work Parlour (1998), Foucault, Kadan and Kandel (2003), etc., approach the problem in ways similar to mine.

Part I

A Dynamic Model of Limit Order Markets

2 Introduction to Part I

In most classical models of market microstructure[4] the market maker plays a central role. The main function of the market maker is to provide liquidity (immediacy) for those who wish to trade, by setting bid and ask quotes. However, there exist many markets, called order-driven, or pure limit order markets, in which there are no market makers (e.g., ECNs, Euronext, Hong Kong, Tokyo, Toronto). In these markets every investor can supply liquidity by placing limit orders in a limit order book.[5] With the recent trend towards larger and more automated exchanges, order-driven markets have become increasingly important.[6] There are also hybrid exchanges (e.g., NYSE, Nasdaq, London), where market makers exist but have to compete with other traders, who supply liquidity by limit orders. In these markets, the number of transactions which involve a market maker is usually small.[7]

The study of liquidity provision when market makers do not exist or only have a limited role is therefore very important in understanding modern financial markets.[8] However, it is quite a complicated problem.[9] To solve it, one would need to know exactly how market prices arise from the interaction of a large number of anonymous traders, who arrive in the market at random times, can choose whether to trade immediately or to wait, and who can

[4] See the survey book by O'Hara (1995).

[5] Limit orders are price-contingent orders to buy (sell) if the price falls below (rises above) a prespecified price. The limit order book is the collection of all price-contingent orders which have not yet been executed.

[6] Nowadays, about half of the world's stock exchanges are organized as order-driven markets, with no designated market makers, while only a few exchanges completely rely on dealer quotes (see Jain (2002)).

[7] For example, on the NYSE the specialists' participation is about 13%, and even less in the case of larger stocks (see Hasbrouck and Sofianos (1993)).

[8] One can argue that market making behavior is important and may arise endogenously even in order-driven markets. However, I know of no data to support this. Anecdotically, there seem to be many traders who submit limit orders (thus providing liquidity), but who do not maintain a continuous presence in the market. Consider the case of a value mutual fund who wishes to acquire a stock based on some analyst reports. If it is willing to wait in the hope of a better execution price in the near future, then it can do so by placing a limit buy order. After the order is executed, the fund might not want to trade again in that stock any time soon. And even if it does, it is not so hard to accept that the next decision to trade in that particular stock can be thought of as being made by a different fund.

[9] There has been significant progress in recent years: see for example the models of Foucault et al. (2003), Goettler et al. (2003), Parlour (1998), and also the discussion in Harris (1998). The main difficulty is not in formulating the problem, but in obtaining a tractable model.

behave strategically by changing their orders at any time.

In this Part I propose a model of an order-driven market which reflects all the features mentioned above. The model is tractable and produces sharp implications about (i) the shape of the limit order book at any point in time, and (ii) the evolution in time of the book, and in particular of the bid and ask prices. The model is in line with known empirical facts, such as the hump shape of the limit order book. It can also explain why following a market buy order both the bid and the ask increase, not only the ask.

I consider a continuous-time, infinite-horizon economy where there is only one asset with no dividends. Buyers and sellers arrive to the market randomly. They either buy or sell one unit of the asset, after which they exit the model. I assume that all traders are liquidity traders, in the sense that their impulse to trade is exogenous to the model. However, they are discretionary, because they have a choice over when to trade, and whether to place a market or limit order. After a limit order is placed, it can be canceled and changed at will. The execution of limit orders is subject to the usual price priority rule, and when prices are equal to the first-in-first-out (FIFO) rule. All agents incur waiting costs, i.e., a loss of utility from waiting. Depending on whether they have low or high waiting costs, traders are patient or impatient. All information is common knowledge, including the limit order book (which is the collection of outstanding limit orders).

In equilibrium it turns out, not surprisingly, that patient agents submit in general limit orders, while impatient agents submit market orders. The new limit orders are always placed inside the bid-ask spread[10], until the spread reaches a minimum level. When this happens let us call the book "full." At that point, patient agents either place a market order or start submitting quick (fleeting) limit orders which in effect behave like market orders. This comes theoretically as a result of a game of attrition amongst the buyers or the sellers.

In order to obtain intuition about the limit order book, I study in more detail a particular case of the model, where one considers just one side of the book, e.g., the sell side: with only patient sellers and impatient buyers. Then the solution can be expressed in closed-form.

Also, to discuss price impact and determine the shape of the limit order-book, I allow

[10]This is not unrealistic, the majority of limit orders are spread-improving: see Biais et al. (1995). One can modify this model, so that the population of patient agents is heterogeneous. Then more patient agents may then place orders away from the market. The model is however more difficult to solve.

for multi-unit market orders, even with very small probabilities. This assumption fixes the levels on which agents place their limit orders, which will then be different for each agent.[11] I show that if the such orders arrive with probabilities which do not decrease too fast (or rather that the agents do not believe so), then the book exhibits a hump shape, i.e., the limit orders will cluster away from the bid and the ask (cf. Biais et al. (1995), and Bouchaud et al. (2002)).

The case when there are no impatient agents can also be solved in closed-form. The resulting equilibrium is quite interesting: buyers and sellers cannot coexist in the limit order book. There are either a lot of sellers in the book and the buyers place market orders, or vice versa. This shows that the existence of impatient agents is important in order for a limit order book to function properly.

I also derive the equilibrium processes that bid and ask prices follow, and the expected time-to-execution for limit orders.[12] The point where the limit order book is full coincides with the time when the bid-ask spread is minimum (an interesting fact since the tick size is zero). One may call this minimum spread the competitive bid-ask spread. Then, as in Foucault et al. (2003), I define resiliency as the speed with which the bid-ask spread reverts to its competitive level. I then recover their results: the resiliency of the limit order book increases with the proportion of patient traders, but decreases with order arrival rate, etc.

An interesting empirical implication is that after a market sell order both the bid and ask prices decrease, with the bid decreasing more than the ask.[13] As a result, the spread itself widens. The latter fact was obtained by Foucault et al., but since they do not allow for cancellation of limit orders, did not also obtain a decrease in the ask price. The fact that the ask also decreases is documented for example in Biais et al. (1995), who also propose an information explanation.[14]

[11]The fact that agents limit trade at different levels comes from the fact that the FIFO rule breaks the symmetry of the payoffs, which forces orders to have different times to execution.

[12]In this model, since agents can switch places in the book, they can increase expected time-to-execution while keeping utility constant (by getting a higher expected execution price). However, if one assumes that traders preserve the relative order in which they arrived in the book, one can define a meaningful notion of time-to-execution.

[13]The bid decreases mechanically, because a limit buy order is cleared from the book. The fact that the ask decreases reflects the sellers' realization that their reservation value (the bid) has decreased, so they also adjust the ask.

[14]This implication can be generated also in a model with price discovery, such as in Glosten and Milgrom (1985). However, I show that even in the absence of information such dynamics may occur.

The limit order book was analyzed in a variety of ways. The information models, which consider market makers interacting with informed agents, are all static: see Glosten (1994), Chakravarty and Holden (1995), Rock (1996), Seppi (1997) and Parlour and Seppi (2001). Moreover, traders are restricted to placing limit orders, so they do not have a choice to submit market orders. Dynamic models, without market makers, are studied by Parlour (1998), Foucault (1999), Foucault, Kadan and Kandel (2003), Goettler, Parlour and Rajan (2003). However, these models are typically not very tractable, and do not allow for strategic cancellation of limit orders.

Although the present work was developed independently from this literature, it turns out that it is closely related to the work of Foucault, Kadan and Kandel (2003). In their model, waiting costs are also the driving force. An important feature of their model is the existence of discrete prices. This allows them to make a comparison among various tick sizes. However, discrete prices make their model more complicated, so in order to simplify it they have to impose strong assumptions such as: (i) there is no cancellation of limit orders (so agents only make only one decision, when they arrive); (ii) a buyer must always arrive after a seller, and vice versa ; and (iii) new orders have to be improving the existing limit orders by at least a tick. Assumptions (ii) and (iii) make the spread the only state variable, which allows a recursive structure for the solution. As a consequence, their model only focuses on the bid-ask spread, and not on the evolution of the actual bid and ask prices.

An interesting related literature is on liquidity and search costs (e.g., Duffie, Garleanu and Pedersen (2001), Vayanos and Wang (2003)), where buyers and sellers have to search for counter-parties to trade. My contention is that on organized exchanges search costs become one-dimensional and can be better thought of as waiting costs, which leads to the current model.[15] This work is also related to the burgeoning field of econophysics (see Farmer et al. (2003) or Gabaix et al. (2003)). See also the literature on double auctions and bid-ask markets, e.g., Wilson (1986).

[15]Of course, one could still argue that search costs are important, especially when stocks are traded on more than one exchange, or when there is more than one specialist. But, more importantly, search costs become significant when dealing with large (block) trades. Since block trades are not dealt with in my model, it would be interesting if one could merge the two frameworks and shed more light on this issue.

3 The Model

3.1 The Market

I consider a market in an asset which yields no dividends. The buy and sell prices for this asset are determined as the bid and ask prices resulting from trading based on the rules given below. There is a constant range $A > B$ where the prices lie at all times. More specifically, there is an infinite supply when price is A, provided by agents outside the model. Similarly, there is an infinite demand for the asset when price is B. Prices can take any value in this range, i.e., the tick size is zero.

Trading. The time horizon is infinite, and trading in the asset takes place in continuous time. The only types of trades allowed are market orders and limit orders. The limit orders are subject to the usual price priority rule; when prices are equal, the first-in-first-out (FIFO) rule is applied. If several market orders are submitted at the same time, only one of them is executed, at random, while the other orders are canceled.[16] Limit orders can be canceled for no cost at any time. There is also no delay in trading, both types of orders being posted or executed instantaneously. Trading is based on a publicly observable limit order book,[17] which collects all the limit orders that have not been canceled or executed.

Agents. The market is populated by traders who arrive randomly to the market, and choose strategically between market and limit orders. They are liquidity traders, in the sense that they want to trade the asset for reasons exogenous to the model. The traders are either buyers or sellers; their type is fixed from the beginning and cannot change. Buyers and sellers trade at most one unit, after which exit the model forever.

Traders are risk-neutral, so their instantaneous utility function (felicity) is linear with

[16] To justify this assumption, it is best to think of a market buy (sell) order as a limit order with limit price equal to the ask (bid). Then if several market orders are submitted at the same time, one of them is randomly executed, while the others remains as limit orders, which can be freely canceled.

[17] In recent years, limit order books have become increasingly available to traders. For example, in NYSE the OpenBook system (introduced in October 2000) allows traders—for a monthly fee of $50/month per screen—to see all limit orders in real time (with a 5-second delay). However, orders from the trading floor, or stop-loss orders are not visible. Another example is Nasdaq Level II, which displays the best bids and offers from market makers and ECNs, and which is publicly available to registered traders. However, other limit orders from the market makers are not available on Nasdaq, except through a premium system like PowerView or DepthView, and even there the visible depth is limited to the best five quotes. For a final example, in Euronext traders can see the best five quotes on each side, while Euronext members have access to the whole limit order book, except hidden quantities and ID codes.

price.[18] By convention, felicity is equal to price for sellers, and minus the price for sellers. Traders discount the future in a way proportional to the expected waiting time. Thus, if τ is the random execution time and P_τ is the price obtained at τ, the utility of a seller is

$$f_t = \mathsf{E}_t\{P_\tau - r(\tau - t)\},$$

Similarly, the utility of a buyer is $-g_t = \mathsf{E}_t\{-P_\tau - r(\tau - t)\}$, where I introduce the notation

$$g_t = \mathsf{E}_t\{P_\tau + r(\tau - t)\}.$$

I call f_t the value function of the seller at t, and g_t the value function of the buyer, although in fact g_t equals minus the utility of a buyer.

The discount coefficient r is constant, and can take two values: if it is low, the corresponding traders are called *patient*, otherwise they are *impatient*. Agents' types are determined from the beginning and cannot change.

For simplicity, I assume that the impatient agents always submit market orders. This is not necessary to make the model work, but it simplifies the presentation. In the Appendix, I discuss conditions for the coefficient r such that in equilibrium impatient agents always submit market orders. Assuming this, from now on r denotes only the time discount coefficient of the patient agents.

Arrivals. The four types of traders (patient buyers, patient sellers, impatient buyers, and impatient sellers) arrive to the market according to independent Poisson processes with constant arrival intensity rates

$$\lambda_{PB}, \lambda_{PS}, \lambda_{IB}, \lambda_{IS}.$$

By definition, Poisson arrival with intensity λ implies that the number of arrivals in any interval of length T has a Poisson distribution with parameter λT. The inter-arrival times of a Poisson process are distributed as an exponential variable with the same parameter λ. The mean time until the next arrival is then $1/\lambda$. The interval until the next arrival is called a *period*.

[18] This model also works with exponential time discounting, but the resulting formulas are more complicated.

Strategies. Because of Poisson arrivals, the game must be set in continuous time. Since there is no universally accepted standard of continuous time game theory, I define in Part II the game theoretic setup that seems the most appropriate to our case: stochastic multi-stage game with observed actions, by extending a framework created by Bergin and MacLeod (1993). All information, together with agents' strategies and beliefs are common knowledge.

One special feature of this model that has to be addressed comes from market orders. Suppose at time t an agent submits a market order. Then the agent exits the model, and the next stage of the game will be played with fewer players. But at which time will this next game be played? No $t + \varepsilon > t$ is satisfactory, because it would imply waiting for a positive time, during which agents lose utility. The best solution, as in auction theory, is to "stop the clock." Then the next game is also played at time t, and the clock is restarted only when in the stage game no agent submits a market order. Allowing for clock stopping in continuous time game theory requires some care, and it is done in Part II.

An important benefit of setting the game in continuous time is that agents can respond immediately. More precisely, one can use strategies that specify: "Keep the limit order at a_1 as long as the other agent stays at a_2 or below. If at some time t the other agent places an order above a_2, then *immediately after* t undercut at a_2." In the rest of this Part, I use this type of strategies freely.

The notions of equilibrium used are sub-game perfect equilibrium, and Markov perfect equilibrium (see Fudenberg and Tirole, ch. 13). One other notion that is important in this framework is that of *competitive* equilibrium, which is a sub-game perfect equilibrium such that at each time all buyers have the same value function (and similarly for sellers). These are discussed in more detail in Part II.

3.2 Discussion.

I now discuss some of the features of this model. One strong assumption is that prices lie within a range $[B, A]$, and that A and B are known by everybody with certainty. Clearly, a more realistic assumption would be to make A and B random, or even stochastic, perhaps as prices coming from valuations of informed traders. In fact, one can think of A and B as summarizing information about the asset, while within this range prices fluctuate due

to the exogenously specified order flow (of course, one should not artificially separate order flow from information). This interpretation would imply that empirical implications of this model would be more believable when obtained in high frequencies, when less information is likely to arrive.

One should be clear that this is not an asymmetric information model. As pointed out by Foucault et al. (2003), in this kind of models, frictions such as the bid-ask spread are completely determined by (i) the waiting costs of agents, and (ii) strategic rent-seeking by patient traders. But this is not so unrealistic. Huang and Stoll (1997) for example estimate that on average approximately 90% of the bid-ask spread is due to non-informational frictions ("order-processing costs").

Another strong assumption is having independent Poisson arrivals. From a qualitative standpoint, more important than the actual Poisson distribution is the independence of the increments of the arrival process. This implies that at each point during the period between two successive arrivals, the agents face the same market conditions, which allows for a relatively simple description of the equilibrium.

4 Equilibrium: One Side of the Book

In this section, I analyze the sell-side of the limit order book, by assuming only two types of traders: patient sellers and impatient buyers. With the notation given above, $\lambda_{PB} = \lambda_{IS} = 0$. (By symmetry, one can derive similar results for the buy-side.) This case proves to be quite tractable. Moreover, it is also useful for understanding the general case, which can be thought as merging two one-sided models.

4.1 Main Intuition and Discussion

Here is some intuition for the equilibrium. Imagine the limit order book is empty, and a patient seller (labeled "1") arrives first to this market. Then, until some other agent arrives, trader 1 optimally behaves like a monopolist, and submits a limit sell order at $a_1 = A$.[19] Suppose now a second patient seller (labeled "2") arrives. Now both sellers compete for the

[19] I have assumed implicitly that if the only limit sell orders in the book are at A, a market order first clears the orders in the book, and only after relies on the infinite supply at A.

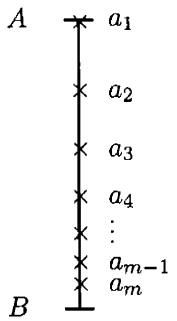


Figure 1: The limit order book with m sellers.

possibility of a market order from incoming impatient buyers. If trader 1 could not cancel his limit order at A, then trader 2 would surely place her limit order at $a_2 = A - \delta$, for some very small δ. But trader 1 can change his limit order, so a price war would likely follow. In order for both traders to be satisfied, trader 2 needs to place her limit order at a level $a_2 < A$ low enough so that trader 1 would be indifferent between keeping his limit order at a_1 (thus having a chance to become a monopolist again after trader 2's order gets cleared), and switching places with trader 2. Now, being indifferent in this case exactly means that both traders get the same expected utility from their strategies. Recall that when traders derive the same expected utility (value) from their equilibrium strategies the equilibrium is called competitive.[20] Of course, the values a_1 and a_2 are determined in equilibrium, and depend on what other agents will do: imagine that instead of an impatient buyer who places a market order at a_2, a patient seller comes, who will place a limit order at a_3, etc.

To summarize the above discussion, when there are m sellers in the book, in equilibrium it must be that some trader has a limit order at the ask a_m. As long as the orders of others are above a_m, their exact levels do not matter (this is because the incoming market orders are one-unit). However, if the trader at the ask tries to place the order higher, then some other agent immediately will jump at a_m.

[20]One can imagine a different scenario, when all patient sellers queue their limit orders at A until the expected utility of the last trader equals the reservation value B. How can such a non-competitive equilibrium be sustained? By Nash threats: Trader 1 can threaten with competitive behavior if trader 2 does not queue behind him at A. Trader 2 is better off by complying as long as she expects trader 3 to do the same and queue behind her. Although these other, non-competitive, equilibria are interesting in their own right, I focus on competitive equilibria since they are the more likely outcome of large, anonymous order-driven markets.

General Discussion

One might wonder what happens if one relaxes the assumption that agents can only trade *one* unit of the asset. Then it seems that trader 2 could do better by buying trader 1 out and becoming a monopolist until the arrival of a new agent.[21] In some sense, trader 2 could profit from hoarding liquidity and thus becoming an endogenous market maker.[22]

The behavior of an endogenous market maker (multi-unit supplier of liquidity) is beyond the scope of this paper, and in order to keep the model tractable, I will simply not allow it. However, since I prefer to think that the most important intuitions of this paper are not model specific, some discussion is in order.

First of all, it is not straightforward how to model the behavior of a multi-unit supplier of liquidity. In order for some trader to hoard liquidity in the way mentioned above, the following conditions have to be satisfied: there are no constraints on borrowing or short-selling; the trader is risk-neutral (so there are no inventory issues); and the arrival rates are independent of the bid-ask spread (otherwise, if the market maker tries to extract too high a rent, then competing liquidity suppliers will arrive with a much higher probability). These conditions are clearly unrealistic, so it is safe to assume that there are limits to endogenous market making. Then an extension of my model where patient traders supply at most n units of the asset (with n a low number) seems quite realistic.

Also, allowing endogenous market making raises a more serious problem. Classical market microstructure, which has focused mainly on the market maker, has been in some sense forced to accept that only the arrival of new information (or at least the possibility of new information) could move prices. Otherwise, why would the arrival of a market order move the quotes? If new information did not change the expected value of the asset, then presumably the market maker could just use inventory to replace the liquidity that has been consumed, without changing quotes. This goes to the heart of the issue, and shows why modeling the market maker is a difficult task.

There are two reasons why the market maker might move prices even in the absence

[21] More precisely, trader 2 can place a limit buy order at some level slightly above the expected utility of trader 1. Then trader 1 would be better off accepting the offer immediately, and trader 2 would become a monopolist instead of competing, so she would also be better off.

[22] See Bloomfield, O'Hara and Saar (2003) for evidence of endogenous market making in experimental order-driven markets.

of information: First, since the order flow is usually positively auto-correlated (a feature not present in this paper), a profit maximizing market maker might decide to change prices to take advantage of the likelihood that the next market order goes in the same direction. Second, and more importantly, is the market maker only a profit maximizer? To some extent, a market maker fulfills also the role of an (inter-temporal) Walrasian auctioneer. Such an auctioneer, when a demand imbalance appears (proxied by a market order in a continuous-trading environment), would adjust market clearing prices (proxied by quotes) in the direction of the imbalance.

4.2 Equal Arrival Rates

I assume a dynamic market clearing condition, namely that the arrival rates of patient sellers and impatient buyers are equal:

$$\lambda = \lambda_{PS} = \lambda_{IB} > 0.$$

Start with a competitive stationary Markov equilibrium. The idea is to find necessary conditions for such an equilibrium to exist, and then to prove that the conditions are also sufficient. As mentioned in the previous section, the Markov equilibrium is defined with respect to the state variable given by the number of sellers m in the limit order book. Since in a competitive equilibrium all agents have the same utility, denote by f_m the utility of an agent in state m.

I enumerate a few results about the equilibrium that will be proved in Appendix A. First, the number of states must be finite. Denote by M the largest state. As m increases, each seller is strictly worse off. In state M it must be true that $f_M = B$, otherwise another patient seller would be tempted to join in. In this state the bottom seller has a mixed strategy: at the first arrival in an independent Poisson process with intensity μ, the seller will place a market order and exit.[23] Observe that from a state $m = 1, \ldots, M-1$, the system can go to one of two states:

- $m-1$, if an impatient buyer arrives—after random time T_1;

[23]There is another possibility for a mixed strategy, but it leads to the same formulas. See Proposition 31.

- $m+1$, if a patient seller arrives—after random time T_2.

Inter-arrival times of Poisson processes are exponentially distributed, so the arrival of the first of the two states happens at $\min(T_1, T_2)$, which is exponential with intensity 2λ. Then each event happens with probability $1/2$. One obtains the formula $f_m = \frac{1}{2}(f_{m-1} + f_{m+1}) - r \cdot \frac{1}{2\lambda}$. Denote by

$$\varepsilon = \frac{r}{\lambda}.$$

The formula becomes

$$2f_m + \varepsilon = f_{m-1} + f_{m+1}.$$

From the terminal state M, the system can go to

- $M-1$, if an impatient buyer arrives—after random time T_1;
- M, if a patient seller arrives—after random time T_2;
- $M-1$, if a current seller places a market order and exits—after random time T_3.

One can ignore the arrival of a new patient seller, since it does not affect the state (the seller places a market order at B and exits). Then one gets the formula $f_M = f_{M-1} - r \cdot \frac{1}{\lambda+\mu}$. Define

$$u = \frac{\lambda}{\lambda+\mu} \in (0,1].$$

The formula becomes

$$f_M + u\varepsilon = f_{M-1}.$$

Define also (this will be justified later):

$$f_0 = A.$$

All these equations can be put together in the following definition.

Definition 1. *Let $u \in (0,1]$ and $M > 0$ an integer. A sequence f_m, $m = 1, \ldots, M$, is called*

a (u, M)-chain if

$$\begin{cases} f_0 = A, \\ 2f_m + \varepsilon = f_{m+1} + f_{m-1}, \quad m = 1, \ldots, M-1 \\ f_M + u\varepsilon = f_{M-1}. \end{cases} \tag{1}$$

The (u, M)-chain f_m is called maximal if $f_M = B$.

An important intuition for a (u, M)-chain is that it f_m decreases for $m < M$, after which it starts increasing.

Proposition 1. *Suppose f_m is extended above M via the middle equation of (1). Then if $i < M$ is an integer, not necessarily positive, the following recursive relation holds:*

$$f_{M-i-1} = f_{M-i} + \varepsilon(i + u).$$

In particular, a (u, M)-chain f_m is strictly decreasing in m if $m < M$ and strictly increasing if $m > M$.

What was done so far was to show that the value functions for any competitive stationary Markov equilibrium form a maximal (u, M)-chain. The next result shows that given A, B and ε, such a chain is unique and gives explicit formulas to calculate it.

Proposition 2. *Given $A > B$ and $\varepsilon > 0$, there exists a unique maximal (u, M)-chain. Any sequence f_m that forms a maximal (u, M)-chain satisfies the formula*

$$f_m = A - bm + \frac{\varepsilon}{2} m^2, \tag{2}$$

where

$$b = \varepsilon \sqrt{(\tfrac{1}{2} - u)^2 + \frac{2(A-B)}{\varepsilon}} = \varepsilon(M - \tfrac{1}{2} + u). \tag{3}$$

The integer $M > 0$ and real number $u \in (0, 1]$ are the unique ones for which

$$M = \frac{1}{2} - u + \sqrt{(\frac{1}{2} - u)^2 + \frac{2(A-B)}{\varepsilon}} \tag{4}$$

is an integer.

Proof. See Appendix A. □

I now state the main result of this Section.

Theorem 3. *Given A, B, r, λ, there exists a unique competitive stationary Markov equilibrium of the game. Let $\varepsilon = \frac{r}{\lambda}$, and M, u, f_m as in Proposition 2. Then in equilibrium there are at most M limit orders in the book, and the ask price in state $m = 1, \ldots, M$ is given by*

$$a_m = f_{m-1}, \quad \text{if } m < M; \tag{5}$$

$$a_M = B + \varepsilon. \tag{6}$$

The value function in state m is given by f_m. The strategy of each agent in state m is the following:

- *If $m = 1$, then place a limit order at $a_1 = A$.*
- *If $m = 2, \ldots, M-1$, place a limit order at any level above a_m, as long as someone stays at a_m (or below). If not, then place an order at a_m.*
- *If $m = M$, the strategy is the same as for $m = 2, \ldots, M-1$, except for the bottom seller at a_M, who exits (by placing a market order at B) after the first arrival in a Poisson process with intensity $\mu = \frac{\lambda}{u} - \lambda$.*
- *If $m > M$, then immediately place a market order at B.*

Proof. See Appendix A and the discussion below. □

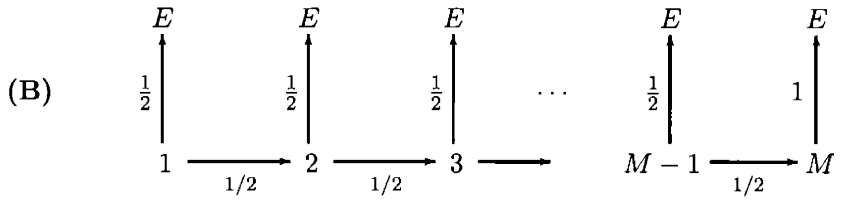
Notice that there is some ambiguity in the way strategies are formulated. They can support several equilibria, all of them payoff-equivalent. For example, in one equilibrium, in state m each agent places an order at a_m, regardless of what the others are doing. Another equilibrium is the one in which no agent changes the order until execution. Then in state m, the limit order book has limit orders at $a_1, \ldots, a_m$. This is the equilibrium suggested at the beginning of this section.

This ambiguity arises from the fact that market orders are for only one unit, so in each state only the limit order at the ask matters. In fact, there is ambiguity also in the identity

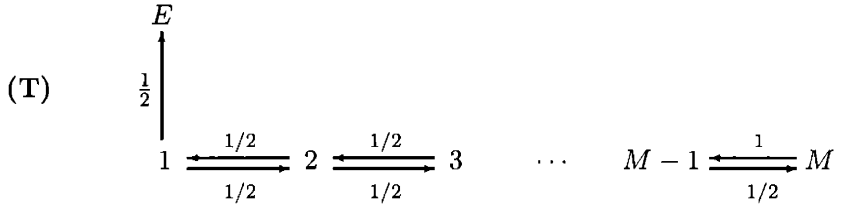
of the players. As mentioned above, it does not matter who has the order at the ask as long as there is one. That divides the agents essentially into two types:

- The **Bottom** agent, who has the lowest offer in the book, placed at a_m.
- The **Top** $m-1$ agents, who are placed above (or at a_m after the Bottom agent, due to the FIFO rule).

Here is the state diagram for the Bottom seller, where E represents exit for this agent (when an impatient seller comes and clears the order at the ask):



And here is the state diagram for the Top sellers:



(Notice that in state 1 there is no distinction between the Top and Bottom agents.) One obtains then the equations

$$(B) \quad 2f_m + \varepsilon = a_m + f_{m+1}, \quad \text{if } m = 1, \ldots, M-1;$$
$$(T) \quad 2f_m + \varepsilon = f_{m-1} + f_{m+1}, \quad \text{if } m = 2, \ldots, M-1.$$

Since the equilibrium is competitive, by definition in state m the value functions are all equal, so it follows that $a_m = f_{m-1}$ for all $m = 2, \ldots, M-1$. Since in state 1 the seller is a monopolist, $a_1 = A$. This justifies the definition: $f_0 = A$.

Now it is clear why the strategies defined by Theorem 3 give a sub-game perfect equilibrium: In each state m no agent wants to deviate. The bottom trader will not go lower than a_m, because then he will lose in expectation. If he tries to go above a_m, then some top agent will immediately become the bottom one.

One can also calculate

$$\begin{aligned} a_M &= B + \varepsilon, \\ a_{M-1} &= B + \varepsilon + 2u\varepsilon, \\ a_{M-2} &= B + 3\varepsilon + 3u\varepsilon, \\ a_{M-3} &= B + 6\varepsilon + 4u\varepsilon, \\ &\vdots \\ a_1 &= A. \end{aligned}$$

It is interesting to see how f_m depends on m and B:

$$\frac{\partial f_m}{\partial m} = -\varepsilon \left((M - \tfrac{1}{2} + u) - m \right); \tag{7}$$

$$\frac{\partial f_m}{\partial B} = \frac{m}{M - \frac{1}{2} + u}. \tag{8}$$

(Notice that $\frac{\partial f_m}{\partial B} < 1$ when $m < M$.) These formulas are important to get intuition about the equilibrium limit order book with both sellers and buyers (then in some sense B represents the bid price). Here is a heuristic argument: suppose the system is in an equilibrium with m sellers and $n+1$ buyers. Then the bid price $B = g_{m,n}$. One calculates

$$\frac{\partial f_{m,n}}{\partial n} = \frac{\partial f_{m,n}}{\partial B} \cdot \frac{\partial g_{m,n}}{\partial n} < \frac{\partial g_{m,n}}{\partial n}. \tag{9}$$

This implies that when a new buyer arrives, the sellers are better off by less than the buyers are worse off (recall that $g_{m,n}$ is minus the buyers' utility). Another, more rigorous argument, will be given in the section that solves for the equilibrium in the general case.

4.3 Different Arrival Rates

In this section, I assume that the patient sellers arrive faster than the impatient buyers:

$$\lambda_1 = \lambda_{PS} > \lambda_2 = \lambda_{IB}.$$

As in the case of equal arrival rates, one starts with a competitive stationary Markov equilibrium, and looks for for necessary conditions. Denote by f_m is the value function of a seller in the state where there are m sellers in the book. From the results proved in Appendix A, the number of states is finite, so there exists a largest state M. Moreover, $f_M = B$. From the state $m = 1, \ldots, M-1$ the the system can go to one of two states:

- $m+1$, if a patient seller arrives—after random time $T_1 \sim \exp(\lambda_1)$;
- $m-1$, if an impatient buyer arrives—after random time $T_2 \sim \exp(\lambda_2)$.

The arrival of the first of these two states happens at $\min(T_1, T_2) \sim \exp(\lambda_1 + \lambda_2)$. Then the first event happens with probability

$$\omega = \frac{\lambda_1}{\lambda_1 + \lambda_2} > \frac{1}{2},$$

and the second with probability $1 - \omega$. Define

$$\varepsilon = \frac{r}{\lambda_1 + \lambda_2}.$$

One then gets the equation

$$f_m + \varepsilon = \omega f_{m+1} + (1-\omega) f_{m-1}.$$

The characteristic equation $\omega x^2 - x + (1-\omega) = 0$ has two roots: $x_1 = 1$ and $x_2 = \alpha$, where

$$\alpha = \frac{1-\omega}{\omega} < 1.$$

Imposing the condition $f_0 = A$ as in the equal arrival case, the general solution is

$$f_m = A - C(1-\alpha^m) + \varepsilon\beta m, \quad \text{with } C \text{ arbitrary,} \tag{10}$$

where

$$\beta = \frac{1}{2\omega - 1}.$$

The constant C is determined by looking at the final state M. In this state, suppose the bottom agent has the mixed strategy to place a market order at the first arrival in a Poisson process with intensity μ. Arrival of patient seller does not matter, so one gets the equation $(\lambda_2 + \mu)f_M + \varepsilon = (\lambda_2 + \mu)f_{M-1}$. Define

$$s = \frac{\mu}{\lambda_1 + \lambda_2}, \qquad u = \frac{1}{1 - \omega + s}.$$

Then one gets

$$f_M + u\varepsilon = f_{M-1}.$$

As in the case of equal arrival rates, one can define a (u, M)-chain.

Definition 2. *Let $u \in (0, \frac{1}{1-\omega}]$ and $M > 0$ an integer. A sequence f_m, $m = 1, \ldots, M$, is called a (u, M)-chain if*

$$\begin{cases} f_0 = A, \\ f_m + \varepsilon = \omega f_{m+1} + (1 - \omega) f_{m-1}, \quad m = 1, \ldots, M - 1 \\ f_M + u\varepsilon = f_{M-1}. \end{cases} \tag{11}$$

The (u, M)-chain f_m is called maximal if $f_M = B$.

As in the previous Section, a (u, M)-chain decreases for $m < M$ and increases for $m > M$.

Proposition 4. *Suppose f_m is extended above M via the middle equation of (11). Then if $i < M$ is an integer, not necessarily positive, the following recursive relation holds:*

$$f_{M-i-1} = f_{M-i} + \varepsilon\big((\beta + u)(\alpha^{-i} - 1) + u\big).$$

In particular, a (u, M)-chain f_m is strictly decreasing in m if $m < M$ and strictly increasing if $m > M$.

One can also prove the following result.

Proposition 5. *Given $A > B$, $\omega > \frac{1}{2}$ and $\varepsilon > 0$, there exists a unique maximal (u, M)-chain. Any sequence f_m that forms a maximal (u, M)-chain satisfies the formula*

$$f_m = A - C_M(1 - \alpha^m) + \varepsilon\beta m, \tag{12}$$

where

$$C_M = \frac{\varepsilon(\beta + u)}{\alpha^{M-1} - \alpha^M}. \tag{13}$$

The integer $M > 0$ and real number $u \in (0, \frac{1}{1-\omega}]$ are the unique ones for which

$$A - B + \varepsilon\beta M = \frac{\varepsilon(\beta + u)}{\alpha^{M-1} - \alpha^M}(1 - \alpha^M). \tag{14}$$

An interesting feature of these formulas when $\omega > \frac{1}{2}$ is that one can calculate the limit of equation (12) when $\varepsilon \to 0$. To see why, notice that $f_M = B$, so $C_M(1-\alpha^M) = A-B+\varepsilon\beta M$. The number α^M is of the order of ε, so M is of the order of $\log\frac{1}{\varepsilon}$. This implies that $C_M = A - B$ modulo terms of order $\varepsilon \log\frac{1}{\varepsilon}$, which is of smaller order for example than $\varepsilon^{1/2}$. In the end, one gets the following formula

$$f_m \approx B + (A - B)\alpha^m \quad \text{if } \varepsilon \approx 0. \tag{15}$$

I now come back to the description of the equilibrium.

Theorem 6. *Given $A, B, r, \lambda_1 > \lambda_2$, there exists a unique competitive stationary Markov equilibrium of the game. Let $\varepsilon, \alpha, \beta$ as above, and M, u, f_m as in Proposition 5. Then in equilibrium there are at most M limit orders in the book, and the ask price in state $m = 1, \ldots, M$ is given by*

$$a_m = f_{m-1}, \quad \text{if } m < M; \tag{16}$$

$$a_M = B + (1 - \omega)\varepsilon. \tag{17}$$

The value function in state m is given by f_m. The strategies of agents are the same as those in Theorem 3.

4.4 Multi-Unit Market Orders

When agents have one-unit demands, the only important limit order in equilibrium is the one at the ask, and the others limit orders can be at any level above the ask. However, assume that multi-unit market orders may arise even with very small probability. To be more precise, let k be the maximum number of units that an order can have, with positive probability. Then in equilibrium I show that the last k levels in the order book are fixed. Assume that patient sellers still arrive with only one unit to sell. Define

$$\begin{cases} \lambda = \text{ arrival rate of patient sellers;} \\ \lambda_i = \text{ arrival rate of } i\text{-unit impatient buyers, } i = 1, \dots, k. \end{cases}$$

Assume that

$$\lambda_i > 0 \text{ for all } i = 1, \dots, k.$$

Moreover, as in the previous section, one wants the sellers to arrive faster than the units demanded by the buyers. This is equivalent to

$$\lambda > \sum_{i=1}^{k} i\,\lambda_i. \tag{18}$$

As before, the number of states is finite, so there exists a largest state M. Moreover, $f_M = B$. From the state $m = 1, \dots, M-1$ the the system can go to one of the following states:

- $m+1$, if a patient seller arrives;
- $m-i$, $i = 1, \dots, k$ if an impatient i-buyer arrives.

In state M there is some randomization: the bottom seller may leave after the first arrival of a Poisson process with intensity μ.

Definition 3. *Let $\mu \geq 0$ and $M > 0$ an integer. A sequence f_m, $m = 1, \dots, M$, is called a*

(μ, M)-chain if

$$\begin{cases} f_0 = f_{-1} = \cdots = f_{1-k} = A, \\ (\lambda + \sum_{i=1}^{k} \lambda_i) f_m + r = \lambda f_{m+1} + \sum_{i=1}^{k} \lambda_i f_{m-i}, \\ (\sum_{i=1}^{k} \lambda_i + \mu) f_M + r = (\lambda_1 + \mu) f_{M-1} + \sum_{i=2}^{k} \lambda_i f_{M-i}. \end{cases} \tag{19}$$

The (μ, M)-chain f_m is called maximal if $f_M = B$.

Now one can put to use equation (18) to show the following result.

Proposition 7. *A maximal (μ, M)-chain f_m is strictly decreasing in m if $m < M$ and strictly increasing if $m > M$. It satisfies the formula*

$$f_m = C_0 + C_1 \alpha_1^m + C_2 \alpha_2^m + \cdots + C_k \alpha_k^m + a\, m, \quad \textit{where} \tag{20}$$

$$a = \frac{r}{\lambda - \sum_{i=1}^{k} i\lambda_i} > 0 \quad \textit{and} \quad |\alpha_1|, |\alpha_2|, \ldots, |\alpha_k| < 1. \tag{21}$$

The complex numbers: $\alpha_0 = 1, \alpha_1, \ldots, \alpha_k$ are the roots of the polynomial

$$P(X) = \lambda X^{k+1} - (\lambda + \sum_{i=1}^{k} \lambda_i) X^k + \sum_{i=1}^{k} \lambda_i X^{k-i}.$$

Proof. Consider the middle equation in (19) for $m = M$ (by extending f_m to be defined for $m = M + 1$). Subtracting that equation from the bottom one, one obtains

$$\mu(f_{M-1} - f_M) = -\lambda(f_M - f_{M+1}).$$

Equation (18) can be used to show that the sequence $\phi_m = f_m - f_{m+1}$ is always decreasing in m. The above equation shows that ϕ_m changes sign at M.

Now the middle equation in (19) is a difference equation that has general solution $f_m = f_m^0 + C_0 \alpha_0^m + C_1 \alpha_1^m + \cdots + C_k \alpha_k^m$, where f_m^0 is a particular solution, and $\alpha_0 = 1, \alpha_1, \ldots, \alpha_k$ are the roots of $P(X)$ (which is the polynomial corresponding to the recursive equation). If one tries a particular solution $f_m^0 = a\, m$, one gets that indeed $a = \frac{r}{\lambda - \sum_{i=1}^{k} i\lambda_i}$. Moreover, if one defines $Q(X) = P(X)/(X - 1)$, the roots of $Q(X)$ are $\alpha_1, \ldots, \alpha_k$. One calculates

$$Q(X) = \lambda X^k - (\lambda_1 + \cdots + \lambda_k) X^{k-1} - \cdots - (\lambda_{k-1} + \lambda_k) X - \lambda_k.$$

For each $i = 1, \ldots, k$, consider the equation $Q(\alpha_i) = 0$, which is the same as $\lambda = (\lambda_1 + \cdots + \lambda_k)\alpha_i^{-1} + \cdots + (\lambda_{k-1} + \lambda_k)\alpha_i^{-(k-1)} + \lambda_k \alpha_i^{-k}$. Suppose $|\alpha_i| \geq 1$. Then using the previous equation one gets $\lambda \leq (\lambda_1 + 2\lambda_2 + \ldots + k\lambda_k) < \lambda$, contradiction. This implies that $|\alpha_i| < 1$. □

The description of the equilibrium is the following:

Theorem 8. *Given A, B, r, λ and $\lambda_i, i = 1, \ldots, k$ which satisfy the inequalities above, there exists a unique competitive stationary Markov equilibrium of the game. Denote by $i_0 = \min\{k, m\}$. In equilibrium there are at most M limit orders in the book, and if $i = 1, \ldots, i_0$, then the level of the i'th limit order (counted from bottom up) in state $m < M$ is given by*

$$a_i(m) = \frac{\lambda_k f_{m-k} + \lambda_{k-1} f_{m-k+1} + \ldots + \lambda_i f_{m-i}}{\lambda_k + \lambda_{k-1} + \ldots + \lambda_i}, \tag{22}$$

where by convention $f_0 = f_{-1} = \cdots = f_{1-k} = A$. The value function in state m is given by f_m. The strategy of each agent in state m is the following:

- *If $m = 1$, then place a limit order at $a_1(1) = A$.*
- *If $m = 2, \ldots, M-1$, look at the bottom k levels (or at all m levels if $m < k$), which are $a_1(m), \ldots, a_{i_0}(m)$. If any of them is not occupied, occupy it. Anything above $a_{i_0}(m)$ does not matter.*
- *If $m = M$, the strategy is the same as for $m = 2, \ldots, M-1$, except for the bottom seller at a_M, who exits (by placing a market order at B) after the first arrival in a Poisson process with intensity μ.*
- *If $m > M$, then immediately place a market order at B.*

One can make these formulas more explicit. There are two cases, depending on whether the k-unit market orders clear all the limit orders in the book or not.

Case 1: $m \geq k$.

$$\begin{aligned}
a_m &= f_0, \\
a_{m-1} &= f_1, \\
&\vdots \\
a_{k+1} &= f_{m-k-1}, \\
a_k &= f_{m-k}, \\
a_{k-1} &= \frac{\lambda_k f_{m-k} + \lambda_{k-1} f_{m-k+1}}{\lambda_k + \lambda_{k-1}}, \\
a_{k-2} &= \frac{\lambda_k f_{m-k} + \lambda_{k-1} f_{m-k+1} + \lambda_{k-2} f_{m-k+2}}{\lambda_k + \lambda_{k-1} + \lambda_{k-2}}, \\
&\vdots \\
a_1 &= \frac{\lambda_k f_{m-k} + \lambda_{k-1} f_{m-k+1} + \cdots + \lambda_1 f_{m-1}}{\lambda_k + \lambda_{k-1} + \cdots + \lambda_1},
\end{aligned}$$

Note that the levels $a_{k+1}, \ldots, a_m$ were chosen by convention, since it does not matter what happens above a_k.

Case 2: $m < k$.

$$\begin{aligned}
a_m &= \frac{\sum_{i=m}^{k} \lambda_i A}{\sum_{i=m}^{k} \lambda_i} = A, \\
a_{m-1} &= \frac{\sum_{i=m}^{k} \lambda_i A + \lambda_{m-1} f_1}{\sum_{i=m}^{k} \lambda_i + \lambda_{m-1}}, \\
&\vdots \\
a_1 &= \frac{\sum_{i=m}^{k} \lambda_i A + \lambda_{m-1} f_1 + \cdots + \lambda_1 f_{m-1}}{\sum_{i=m}^{k} \lambda_i + \lambda_{m-1} + \cdots + \lambda_1}.
\end{aligned}$$

5 Price Impact of Transactions

Having now a tractable model of the limit order book, one can meaningfully talk about price impact in this model. For example, suppose the system is in state m and a buyer submits a market order for one unit. Then the system moves automatically to state $m-1$ (there is one less seller). The ask price therefore moves from a_m to a_{m-1}. If the buyer wants now to purchase another unit, then he must do so at a different price.

5.1 The Shape of the Limit Order Book

In this model, agents stay on different levels in the book even though they are identical. They do this because they are indifferent between staying at one level and switching on another level (as long as the other agents stay on the same levels). The key assumption here is that some market orders may arrive, even with very small probability. Then it may be optimal for agents to cluster away from the ask, to capture the incoming multi-unit market orders. In fact, one can argue that what matters here are the *expectations* that traders have about the arrival rates of the incoming market orders, and not the actual values. Suppose the agents expect that large market orders will arrive. Then they will stay at higher levels than they would normally do. Therefore, one must keep in mind that the values $\lambda_1, \ldots, \lambda_k$ can be interpreted as representing what agents expect about incoming market orders. When one analyzes the shape of the price impact function, it turns out that a crucial factor is how fast the rates λ_i decrease.

To quantify price impact, I use the theoretical results obtained in Section 4.4. Recall that λ_i is the arrival rate of i-unit impatient buyers ($i = 1, \ldots, k$), and $\lambda > \sum_{i=1}^{k} i\lambda_i$ is the arrival rate of patient sellers. To get some intuition, suppose λ_1 is larger than the rest. For example, consider a function $\phi(i)$ which is decreasing in i, and some small value $\lambda_0 > 0$ such that

$$\lambda_1 = 1 \quad \text{and} \quad \lambda_i = \lambda_0 \phi(i), \text{ if } i \geq 2.$$

Then set $\lambda = \sum_{i=1}^{k} i\lambda_i$ (plus some small number, so that one has indeed $\lambda > \sum_{i=1}^{k} i\,\lambda_i$; in fact it can be equal, it does not change the analysis). The problem I want to investigate is to calculate the price impact function for different choices of $\phi(i)$, $i = 2, \ldots, k$. Take for example

$$\phi_1(i) = \frac{1}{i(i+1)}, \qquad \phi_2(i) = \frac{4}{2^i}.$$

The price impact function in state m is defined as the change in the ask price when i units are bought via a market order of i units:

$$Imp(i, m) = a_{i+1}(m) - a_1(m), \quad \text{as a function of } i \text{ (and } m).$$

To calculate the price impact function, one can apply the theoretical results of the

previous section in the following way: Equation (19) shows that f_m satisfies a recursive formula with initial conditions $f_0 = f_{-1} = f_{1-k} = A$. This fixes k coefficients of f_m in equation (20). To fix the last coefficient, one should use the bottom equation in (19). For simplicity, I choose a different method: Let B become a free parameter, and choose another parameter $\delta > 0$, so that

$$f_1 = A - \delta.$$

This fixes all coefficients C_i, so one can determine M as the first m for which f_m starts increasing (according to Proposition 7). Then one determines B by $B = f_M$. So in the analysis that follows and in Figure 2, instead of B, I use the parameter δ.

In Figure 2, I compare the graphs of the price impact function $Imp(i, m)$ for the two functions $\phi_1 = 1/i(i+1)$ and $\phi_2 = 4/2^i$. In both cases the maximum number of sellers is 41 (so $m < 41$), and I display the results when the book has $m = 10, 20, 30, 40$ limit sell orders. All the graphs shown are for $k = 20$, which means that the sellers in the book believe that market buy orders for more than 20 units appear with zero probability. In the first case (for ϕ_1), the sellers believe that the arrival rate λ_i of i-market orders does not decrease very fast in i, while in the second case they believe that the arrival rate decreases exponentially. The top four plots refer to ϕ_1, and the bottom four to ϕ_2.

Notice that in the case of ϕ_1, the price impact function $Imp(i)$ is concave when $i \leq k$ units are purchased (recall that k is the maximum size of a market buy order that sellers in the book expect to occur). So when m itself is less than k, the price impact function is concave everywhere, which is the case of the first two graphs. The intuition for this finding is the following: each seller above the ask up to level k believes that his limit order will be cleared by a market order with a probability which is not too small. Then instead of clustering near the ask, they prefer to take advantage of the large market orders and cluster above the ask. This leads to a concave price impact function. Above the level k or when the probability of a large market order decreases too fast, the price impact function is linear, and even convex. This is the case for all the graphs for ϕ_2, and in the region where $i > k$ for ϕ_1.

Overall, the conclusion seems to be that for smaller orders the price impact function should be mildly concave, and for larger orders it should be mildly convex. This reflects

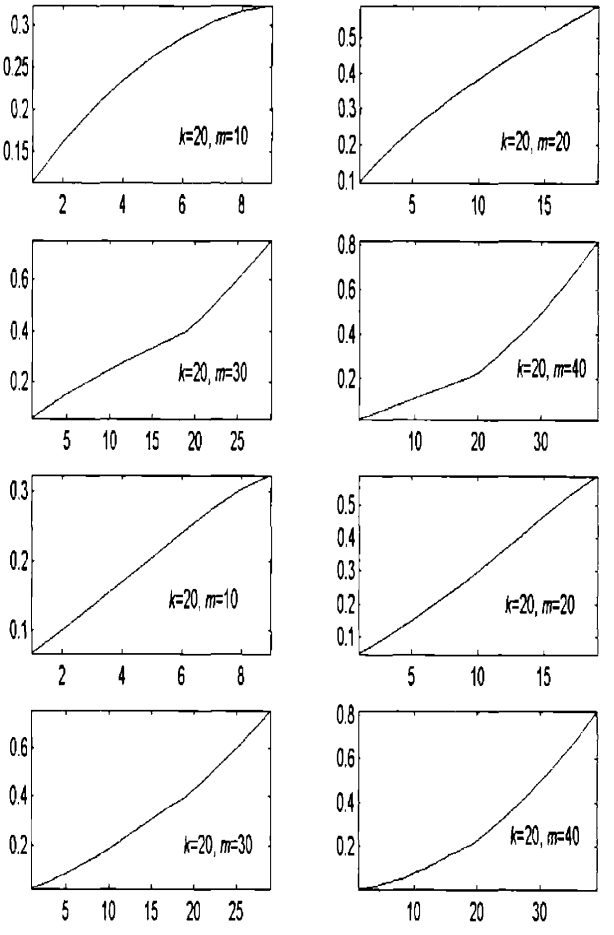


Figure 2: The instantaneous price impact function $Imp(i) = a_{i+1} - a_1$ plotted against i. ($Imp(i)$ is the difference in the level of the $i+1$'st limit sell order above the ask and the ask price.) The values of parameters are $A = 1, r = 0.001, \delta = 0.04, k = 20$, and the arrival rates are $\lambda_1 = 1$; $\lambda_i = \lambda_0 \phi(i), i = 2, \ldots, k$, where $\lambda_0 = 10^{-5}$; $\lambda = \sum_{i=1}^{k} i\lambda_i$. The top four plots correspond to the weight function $\phi(j) = 1/j(j+1)$; the bottom four plots to $\phi(j) = 4/2^j$. Each set of four graphs is considered for the case when there are $m = 10, 20, 30, 40$ sell orders in the book. The number k is the maximum number of units that market buy orders can have, and $\phi(i)$ indicates how fast the arrival rate of i-market orders decreases with i. For the top four plots (when $\phi(i)$ does not decrease too fast), notice the concave shape of the price impact function $Imp(i)$ in the region where $i \leq k$. For the other regions notice either linearity or convexity of price impact.

the existing differences of opinions in the literature, which has not said yet the final word whether the price impact is concave, linear, or convex, and in what range.

5.2 Theoretical vs. Observed Price Impact

In the preceding discussion about price impact, I introduced what might be called "theoretical" price impact, which represents the instantaneous reaction of the ask price to a large market order. When one estimates price impact from data, one should also take into account that agents who submit market orders may strategically want to divide the market order in smaller pieces. Then, the "observed" price impact might be different from the theoretical one.

To understand better this difference, I give an example set in the framework of Section 4.3 (with patient sellers arriving faster than impatient buyers). Suppose the system is in state M, and a strategic agent wants to place a market buy order for two units. Then the first market order for one unit will be placed immediately, and the system moves to $M-1$. Then a dilemma arises: should the agent place a market buy now for the price of a_{M-1}, or wait until the system goes again to state M and place a market buy for $a_M < a_{M-1}$? That depends on the patience of the agent, and on the time the agent has to wait.

For this, one has to calculate the mean expected (first) passage time from state $M-1$ to M: $t_{M-1,M}$. In general, the mean passage time t_{ij} from state i to j can be calculated from the following system of equations:

$$t_{ij} = 1 \cdot P_{ij} + \sum_{k \neq j} (1 + t_{kj}) P_{ik},$$

where P is the Markov transition matrix (with states $0, 1, \ldots, M$):

$$P = \begin{bmatrix} 1-\omega & \omega & 0 & 0 & \cdots & 0 & 0 \\ 1-\omega & 0 & \omega & 0 & \cdots & 0 & 0 \\ 0 & 1-\omega & 0 & \omega & \cdots & 0 & 0 \\ 0 & 0 & 1-\omega & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \omega \\ 0 & 0 & 0 & 0 & \cdots & 1-\omega & \omega \end{bmatrix}. \tag{23}$$

To calculate t_{ij} in general, one looks at the column t^j formed with t_{ij}. Let $\tilde{P}_j$ be the matrix obtained from P by replacing the j'th column by zero. Let e be the column vector formed with ones. Then t^j can be obtained from the formula:

$$t^j = e + \tilde{P}_j \, t^j.$$

In our case $t_{m,M}$ satisfies a recursive formula in m, and one calculates

$$t_{m,M} = \beta(M-m) - \frac{\beta}{\alpha^{-1}-1}(\alpha^m - \alpha^M), \qquad \alpha = \frac{1-\omega}{\omega}, \beta = \frac{1}{2\omega - 1}.$$

So $t_{M-1,M} = \beta - \beta\alpha^M \approx \beta$. Now the mean passage time is calculated in units of time $= \frac{1}{\lambda_1+\lambda_2}$. Denote by r_0 the time discounting coefficient of the agent, and by

$$\varepsilon_0 = \frac{r_0}{\lambda_1 + \lambda_2}.$$

The choice is then between a_{M-1} and $a_M + \varepsilon_0\beta$. If the agent who places market orders can place limit orders instead, then the choice is between f_{M-1} and $f_M + \varepsilon_0\beta$. But $f_{M-1} = f_M + u\varepsilon$. So the agent waits until $t_{M-1,M}$ if and only if $ru > r_0\beta$. I have assumed $s = 0$ for simplicity (so in the state M there is no mixing). Then the agent waits if and only if

$$r_0 < r\frac{2\omega - 1}{1-\omega}.$$

6 Equilibrium: The General Case

For simplicity, I analyze the general case, when all arrival rates are equal

$$\lambda = \lambda_{PB} = \lambda_{PS} = \lambda_{IB} = \lambda_{IS} > 0.$$

Later on, I indicate how similar results can be obtained when the arrival rates are different (the most important case is when $\lambda_1 = \lambda_{PB} = \lambda_{PS}$ is larger than $\lambda_2 = \lambda_{IB} = \lambda_{IS}$).

To get some intuition about the equilibrium, consider a setup similar to that of the one-sided case, but suppose that after after a patient seller (which has a limit sell order at A) a patient buyer arrives to the market. Then the buyer behaves as a monopolist towards the potential incoming impatient sellers, and places a limit buy order at B. In this situation, if the reservation value of the seller is larger than the reservation value of the buyer, they will not be tempted to make offers to one another, and would rather wait to trade with future impatient agents (this happens for most values of the parameters of the model). It follows that patient buyers and sellers behave very much like in the one-sided case, where new patient agents just keep placing bid-ask improving limit orders until it is better to trade immediately rather than wait. Thus, patient agents form two queues, a descending one starting from A, and an ascending one starting from B (see Figure 3 below). However, at some point, the two queues get close to each other, and the patient traders at the bid and the ask may want to trade with each other immediately instead of waiting for impatient agents to place market orders. In this case, I say that the limit order book is full (or saturated). To get a quantitative solution of this problem, one must describe more precisely when the book becomes full.

6.1 Theory

As in the one-sided case, I start by assuming that the system is already in equilibrium, and try to find necessary conditions. This is the difficult part, and at the first reading one may want to skip to the second part of Theorem 9, which is the one most useful for applications. The second part of the Theorem says that, given the solution of some system of partial difference equations in the plane, one can construct a stationary Markov perfect

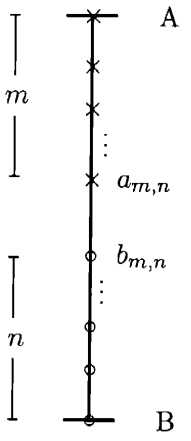


Figure 3: The limit order book with m sellers and n buyers.

equilibrium (MPE) of the game. This allows one to give a lot of examples of an equilibrium limit order book. However, if one wonders in what sense the equilibrium is unique, the following analysis is important.

So assume that the system has reached a competitive, stationary Markov perfect equilibrium. (See Part II for definitions. The Markov structure is given by the pair (m, n) of the number of sellers and buyers.) In state (m, n) at time t, denote by $f_{m,n}$ the expected utility of the sellers and by $g_{m,n}$ the expected utility of the buyers (they do not depend on t because this is a stationary equilibrium). By definition

$$f_{m,n} = \max_{\tau} \mathsf{E}_t\{P^s(\tau) - r(\tau - t)\},$$

where $P^s(\tau)$ is the selling price at the stopping time τ. Similarly,

$$g_{m,n} = \min_{\tau} \mathsf{E}_t\{P^b(\tau) + r(\tau - t)\},$$

where $P^b(\tau)$ is the buying price at τ (recall that $g_{m,n}$ represents minus the utility of the buyers in state (m, n)). Define by $a_{m,n}$ the ask price, and by $b_{m,n}$ the bid price.

Definition 4. *A state is (m, n) is called* **regular** *if in equilibrium traders stay for an expected positive time. Define the* **state region** Ω *as the collection of all regular states.*

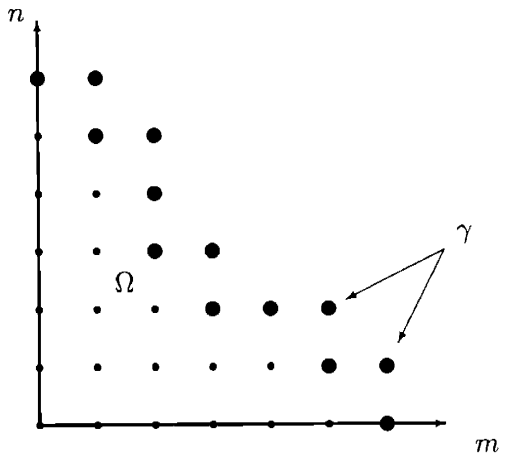


Figure 4: The state region Ω and the boundary γ.

A state in which some agent can exit via a mixed strategy is called **partial**. *If the mixing behavior happens as in Corollary 32, I call the partial state* **rigid** *(this means that mixing is done only by the bottom agent). The* **boundary** γ *of* Ω *is the set of partial states. A state in which the system stays for zero time is called* **fleeting**.

Also, a point in Ω not on γ is called "interior." Proposition 33 implies that all interior points represent states in which agents wait until the arrival of a new agent. As shown in Proposition 33, the partial states only appear when the book becomes full, so partial states will indeed lie on the geometric boundary γ (see Figure 4).

Now Proposition 33 shows that if one starts at each point in Ω and goes along the main diagonal, there is at most one partial state on the boundary γ. If for all starting points in Ω, the corresponding boundary state is partial rigid, I call the equilibrium **rigid**. In words, having a rigid equilibrium means that when the limit order book becomes full, traders always use mixed strategies; moreover, mixing is done in such a way so that all agents have the same value function. (Using Corollary 32, one can show that in a partial rigid state (m, n) the value functions for the buyers and sellers are equal: $f_{m,n} = g_{m,n}$. This in fact is the main reason why I use rigid partial states.)

From now on I assume the system has reached a rigid competitive stationary MPE. Then the resulting state space satisfies Lemma 30, and this implies the existence of a non-

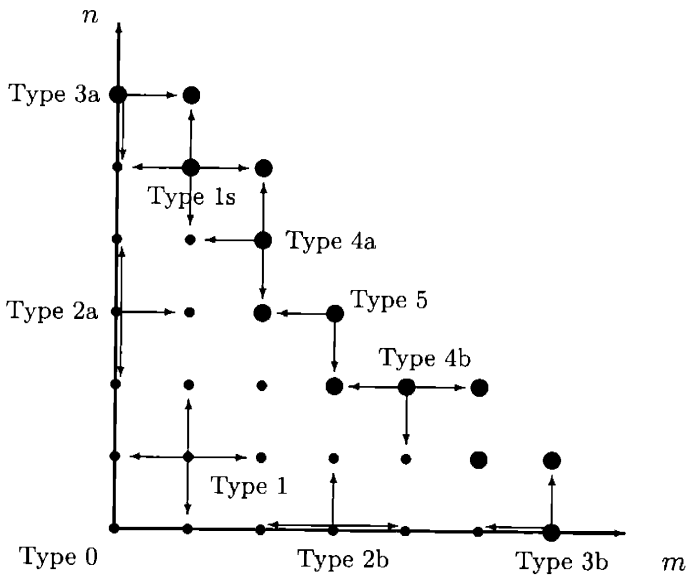


Figure 5: Types of points in the state region Ω.

increasing shape function ϕ that completely characterizes Ω. Using the shape ϕ, one can show that the points in Ω can only be of one of the types described in Figure 5 (use also Corollary 33). (In principle, there is one more type of boundary point, where the shape function ϕ takes the value zero more than once, but it is easy to show that this behavior cannot appear in equilibrium.)

Now, as in the one-sided case, it is a good idea to find a recursive structure for the value functions f and g. From state (m, n) the system can go to the following neighboring states:

- $(m-1, n)$, if an impatient buyer arrives; also, if a patient buyer arrives and submits a fleeting limit order;
- $(m+1, n)$, if a patient seller arrives and submits a (non-fleeting) limit order;
- $(m, n-1)$, if an impatient seller arrives; also, if a patient seller arrives and submits a fleeting limit order;
- $(m, n+1)$, if a patient buyer arrives and submits a limit order;

- $(m-1, n-1)$, if after a positive expected time in state (m, n) a pair of existing patient traders, a buyer and a seller, trade with each other via a fleeting limit order.

From an interior state (m, n) (of Type 1) the system can go only to the states $(m-1, n)$, $(m+1, n)$, $(m, n-1)$, or $(m, n+1)$. The arrival of the first of these four states happens after a random time, which is exponentially distributed with parameter 4λ. Then each event happens with probability $1/4$. One obtains the formula $f_{m,n} = \frac{1}{4}(f_{m-1,n} + f_{m+1,n} + f_{m,n-1} + f_{m,n+1}) - r \cdot \frac{1}{4\lambda}$. Denote again by

$$\varepsilon = \frac{r}{\lambda}.$$

The formula becomes

$$4f_{m,n} + \varepsilon = f_{m-1,n} + f_{m+1,n} + f_{m,n-1} + f_{m,n+1}.$$

This type of reasoning works for all the other states in Ω.

I now describe the equilibrium. The relevant parameters are A, B and $\varepsilon = r/\lambda$, i.e., the ratio between the time discount coefficient and the arrival intensity.

Theorem 9. *Consider a competitive stationary Markov equilibrium of the game with both buyers and sellers, where all types of traders arrive at equal rates. Assume that the equilibrium is also rigid. Then there exists a state region Ω as described in Lemma 30, and drawn in Figure 5. For a state $S = (m, n)$ in Ω, the corresponding value functions f and g satisfy the following equations:*

- *If S is of Type 0,*

$$f_{0,0} = A, \tag{24}$$
$$g_{0,0} = B; \tag{25}$$

- *If S is of Type 1,*

$$4f_{m,n} + \varepsilon = f_{m-1,n} + f_{m+1,n} + f_{m,n-1} + f_{m,n+1}, \tag{26}$$
$$4g_{m,n} - \varepsilon = g_{m-1,n} + g_{m+1,n} + g_{m,n-1} + g_{m,n+1}; \tag{27}$$

- *If S is of Type 1s, then for some $s_{m,n} \geq 0$,*

$$(4 + s_{m,n})f_{m,n} + \varepsilon = f_{m-1,n} + f_{m+1,n} + f_{m,n-1} + f_{m,n+1} + s_{m,n}f_{m-1,n-1}, \tag{28}$$
$$(4 + s_{m,n})g_{m,n} - \varepsilon = g_{m-1,n} + g_{m+1,n} + g_{m,n-1} + g_{m,n+1} + s_{m,n}g_{m-1,n-1}, \tag{29}$$
$$f_{m,n} = g_{m,n}; \tag{30}$$

- *If S is of Type 2a,*

$$f_{0,n} = A, \tag{31}$$
$$3g_{0,n} - \varepsilon = g_{0,n-1} + g_{0,n+1} + g_{1,n}; \tag{32}$$

and similarly for Type 2b.

- *If S is of Type 3a,*

$$f_{0,n} = A, \tag{33}$$
$$(2 + s_{0,n})g_{0,n} - \varepsilon = (1 + s_{0,n})g_{0,n-1} + g_{1,n}; \tag{34}$$

and similarly for Type 3b.

- *If S is of Type 4a, then for some $s_{m,n} \geq 0$,*

$$(4 + s_{m,n})f_{m,n} + \varepsilon = f_{m-1,n} + 2f_{m,n-1} + f_{m,n+1} + s_{m,n}f_{m-1,n-1}, \tag{35}$$
$$(4 + s_{m,n})g_{m,n} - \varepsilon = g_{m-1,n} + 2g_{m,n-1} + g_{m,n+1} + s_{m,n}g_{m-1,n-1}, \tag{36}$$
$$f_{m,n} = g_{m,n}; \tag{37}$$

and similarly for Type 4b.

- *If S is of Type 5, then for some $s_{m,n} \geq 0$,*

$$(4 + s_{m,n})f_{m,n} + \varepsilon = 2f_{m-1,n} + 2f_{m,n-1} + s_{m,n}f_{m-1,n-1}, \tag{38}$$
$$(4 + s_{m,n})g_{m,n} - \varepsilon = 2g_{m-1,n} + 2g_{m,n-1} + s_{m,n}g_{m-1,n-1}, \tag{39}$$
$$f_{m,n} = g_{m,n}; \tag{40}$$

Conversely, for each solution of the system above with $s_{m,n} \geq 0$ and $f_{m,n} \geq g_{m,n}$ there exists a corresponding rigid competitive stationary Markov equilibrium.

Proof. See Appendix A. □

I call the numbers $s_{m,n}$ **slack** variables. They equal $s_{m,n} = \mu_{m,n}/\lambda$, where $\mu_{m,n}$ is the Poisson exit rate in the partial state (m, n) on the boundary. The direct implication of Theorem 9 is useful mainly for questions relating uniqueness of equilibria (so far, it seems that equilibria are very close to each other, or perhaps even unique). However, the more important implication is the converse one. Suppose there is a region Ω of the type mentioned above, such that for all $(m, n) \in \Omega$ there exist numbers $f_{m,n}$, $g_{m,n}$, $s_{m,n}$ which satisfy the equations of Theorem 9. Then the Theorem guarantees the existence of an equilibrium of the game. To understand what kind of strategies give this equilibrium, one needs to define two set of numbers $b_{m,n}$ (bid prices), and $a_{m,n}$ (ask prices) for each $(m, n) \in \Omega$.

Proposition 10. *Consider a competitive stationary Markov equilibrium as in Theorem 9, and let $S = (m, n)$ be a state in Ω. Then the corresponding bid price $b_{m,n}$ and ask price $a_{m,n}$ satisfy*

- *If S is of Type 1,*

$$a_{m,n} = f_{m-1,n}, \tag{41}$$
$$b_{m,n} = g_{m,n-1}; \tag{42}$$

- *If S is of Type 2a,*

$$a_{0,n} = A, \tag{43}$$
$$b_{0,n} = g_{0,n-1}; \tag{44}$$

  and similarly for Type 2b.

- *If S is of Type 5, then for some $s_{m,n} \geq 0$,*

$$a_{m,n} = (1 + s_{m,n}) f_{m-1,n} - s_{m,n} f_{m,n}, \tag{45}$$
$$b_{m,n} = (1 + s_{m,n}) g_{m,n-1} - s_{m,n} g_{m,n}; \tag{46}$$

The formulas for the other types of boundary points are similar.

I now briefly describe the corresponding equilibrium strategies. Let (m, n) be a state which does not necessarily belong to Ω (one must also describe what happens in the fleeting states). Then define the following strategy profile:

- If (m, n) is in Ω but not on the boundary γ, each seller plays the strategy: place a limit sell order at $a_{m,n}$. The seller can also place the limit order at any point above, but in that case some other seller must have a limit order at $a_{m,n}$; otherwise, *immediately* drop to $a_{m,n}$. The strategy works in a similar way for buyers and $b_{m,n}$.

- If (m, n) is in γ, let $\mu_{m,n} = \lambda s_{m,n}$. Then the strategy is similar to the one above, except that the seller at the ask may randomly change the limit order from $a_{m,n}$ to $f_{m,n}$ at a Poisson rate of $\mu_{m,n}$, and the buyer at the bid will immediately accept by placing a market order.

- If (m, n) is a fleeting state with $m, n > 0$, then define the value functions $f_{m,n}$ and $g_{m,n}$ by looking at the values from the corresponding state on γ along the main diagonal. (If the main diagonal does not intersect γ, but one of the coordinate axes, simply

define $f_{m,n} = g_{m,n}$ to be either A or B depending on whether it is the x-axis or the y-axis, respectively.) Then the strategy for each seller is to place a limit order at $f_{m,n}$ and for some buyer to instantly accept it by placing a market order. Which order gets executed is determined randomly.

- If (m, n) is a fleeting state with $n = 0$, define $f_{m,n} = g_{m,n} = B$. Then the strategy for each seller is to place a market order at B (so that only one of them is executed, at random). A similar description is for fleeting states (m, n) with $m = 0$.

Given the equations that all these numbers satisfy, it is not hard to see that the strategy profile just defined is an SPE, and indeed a competitive stationary Markov equilibrium.

Coming back to the expression for the equilibrium in Theorem 9 one can see that both $f_{m,n}$ and $g_{m,n}$ satisfy some finite difference equations, so one may ask if in the limit, when ε is very small, whether f and g do not satisfy some differential equations. It turns out that the answer is yes. Let $\varepsilon = \delta^2$, $x = m\delta$ and $y = n\delta$. Define the functions f and g at the discrete values $(x, y) = (m\delta, n\delta)$ by

$$f(x, y) = f_{m,n}, \qquad g(x, y) = g_{m,n}.$$

Then one gets the following asymptotic result:

Theorem 11. *The solution of the model converges when $\varepsilon = r/\lambda$ is small to the solution of the following system of partial differential equations with a free boundary γ:*

$$\begin{cases} \triangle f = 1, \\ f(0,y) = A, \\ \frac{\partial f}{\partial y}(x,0) = 0, \\ \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} = 0 \text{ at } \gamma; \end{cases} \qquad \begin{cases} \triangle g = -1, \\ g(x,0) = B, \\ \frac{\partial g}{\partial x}(0,y) = 0, \\ \frac{\partial g}{\partial x} + \frac{\partial g}{\partial y} = 0 \text{ at } \gamma; \end{cases} \tag{47}$$

where the free boundary γ is determined by the condition

$$f = g \text{ at } \gamma.$$

The problem has a unique solution, which is symmetric in x and y. The curve γ is slightly concave-down, and passes approximately through the points $(2,0)$ and $(0,2)$.

Proof. See Appendix A. □

This is a Poisson equation in a closed region with mixed-derivative conditions at the boundary. The condition $f = g$ at the boundary determines the free boundary γ, where the limit book is saturated. Since the oblique derivative is never tangent to γ, the problem is well posed, and one can write an algorithm to solve it, using for example finite differences.

6.2 Numeric Results

Having described the equilibrium, one may wonder how to actually find it. This is not a trivial problem: for each value of A, B and ε, one needs to find the state region Ω (or equivalently the boundary γ) and solve the system given in Theorem 9. One could take the brute force approach and for each Ω within reasonable values try to solve the system. For ε small however, the complexity of this approach becomes daunting. The reason is that this is a system of equations with a free boundary γ, and one needs to simultaneously find the solution of the system and the shape of the boundary.

To find the state region Ω, one uses the intuition coming from the asymptotic result in Theorem 11. One sees that the asymptotic boundary is slightly concave, so it is a good idea to try regions state regions Ω which are close to being triangular (bounded by the coordinate

Table 1: Solution in the general case with both buyers and sellers, for $A = 1, B = 0, \varepsilon = 0.09$. Left bottom corner corresponds to state $(0, 0)$. The number in position (m, n) represents the value function $f_{m,n}$ for the sellers in state (m, n). The shape function is $\phi = [0, 1, 1, 2, 2, 2, 3]$, which leads to the state region Ω as in Figure 4. The vector s collects the slack variables along the boundary, starting from $(0, 6)$ down to $(6, 0)$. The value function $g_{m,n}$ for the buyers is given by the formula $g_{m,n} = 1 - f_{n,m}$. The bullets in positions $(3, 4)$ and $(4, 3)$ (which are not in Ω) indicate the departure of the shape of Ω from the triangular one.

1.000	0.965					
1.000	0.905	0.824				
1.000	0.828	0.726	•			
1.000	0.770	0.616	0.500	•		
1.000	0.726	0.526	0.384	0.274	0.176	
1.000	0.697	0.468	0.300	0.177	0.095	0.035
1.000	0.682	0.440	0.260	0.131	0.045	0.000

$$s = [0.21, 3.97, 0.99, 34.34, 2.50, 0.30, 3.47, 0.30, 2.50, 34.34, 0.99, 3.97, 0.21].$$

axes and the line $X + Y = R$). As one increases the size of Ω, one will be forced to take out a few points from the triangle; otherwise, there would be no solution to the system of equations. Indeed, in the examples I computed, the shape of Ω is close to being triangular only when ε is relatively large. As ε gets smaller, points on the diagonal $X + Y = R$ start missing (see Table 1).

I revert now to the more general case when $\lambda_1 = \lambda_{PB} = \lambda_{PS}$ may be different from $\lambda_2 = \lambda_{IB} = \lambda_{IS}$. Define the parameter ω as the percentage of patient traders that arrive to the market:

$$\omega = \frac{\lambda_1}{\lambda_1 + \lambda_2}. \tag{48}$$

So far I only discussed the case $\omega = 1/2$. But the case when $\omega > 1/2$ is even more important. First, it corresponds to empirical evidence (see Biais et al. (1995)), which shows that more

limit orders than market orders arrive to the market.[24] And second, $\omega > 1/2$ has the desirable implication that there is pressure for spreads to revert to small values. If instead impatient agents arrived faster, then there would pressure for spreads to widen towards the value $A - B$.

When patient traders arrive faster than impatient traders, it is reasonable to expect that the waiting costs of patient agents increase. That implies that the limit order book will be more "rarefied" than in the case $\omega = 1/2$. The first guess is that the regions Ω for which one can find solutions are more concave than for $\omega = 1/2$. Indeed this is the case, as can be seen numerically. This is consistent with the limiting case $\omega = 1$ from the next Section, where I show that, if only patient people arrive to the market, the region Ω collapses to the coordinate axes.

7 Equilibrium: The Homogeneous Case

An interesting case to study is when all agents are equally patient ($\lambda_{IS} = \lambda_{IB} = 0$). For simplicity, I also assume that the arrival rates of patient sellers and patient buyers are equal:

$$\lambda = \lambda_{PS} = \lambda_{PB} > 0.$$

It turns out that in this case the limit order book cannot accommodate both buyers and sellers. The reason this happens is because traders lose their incentive to wait. In the one-sided case, sellers were waiting to extract rents from impatient buyers. Now, when all agents are equally patient, they cannot extract rents from each other, so a bargaining game follows. In principle, there can be many equilibria, unless one puts more structure on the bargaining problem. I do not follow this path here. Instead, I give an example of a competitive stationary Markov equilibrium, which seems to be the prototype for all such equilibria. For simplicity, assume $A = 1$, $B = 0$ (this can be done without any loss of generality). Define as before $\varepsilon = r/\lambda$.

As in the general case from the previous Section, one can talk about the state region. In order to describe an equilibrium, one also has to describe what happens in each state

[24] The same empirical research indicates that when the spread is wider the percentage of submitted limit orders increases. However, for technical reasons I do not model spread-dependent arrival intensities.

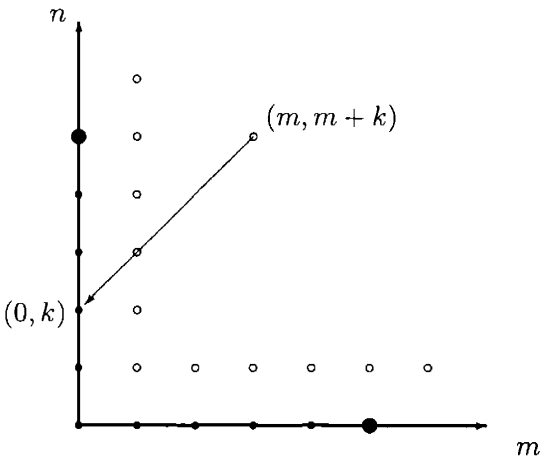


Figure 6: The state region in the homogeneous case. (Surrounding fleeting states are marked with a circle).

(m, n).

Definition 5. *Let M, μ, f_m, a_m be as in the Theorem 3 applied to $A = \frac{1}{2}$ and $B = 0$ (for the bottom half of the limit order book). Define also $g_m = 1 - f_m$ and $b_m = 1 - a_m$, the symmetric values with respect to $1/2$. For $m > M$ extend $f_m = 0$ and $g_m = 1$. Let Ω be the collection of points $(0, m)$ and $(m, 0)$ for $m = 0, \ldots, M$.*

Theorem 12. *Consider any state (m, n) in the plane. Then a competitive stationary Markov equilibrium of the game is given by the following strategy profile:*

- *If $m = 0$ or $n = 0$, apply the same strategy as in the one-sided case. For example, in state $(0, n)$ with $0 < n < M$ at least one buyer places a limit buy order at b_n (and expects that an incoming seller will place a market order).*
- *If $m > n > 0$, place immediately a limit buy (sell) order at f_{m-n} (and expect that some seller (buyer) will immediately accept it).*
- *If $n > m > 0$, place immediately a limit buy (sell) order at g_{n-m}.*
- *If $m = n > 0$, place immediately a limit sell (buy) order at $f_0 = g_0 = 1/2$.*

Note that in each state (m,n) with $m,n>0$, one has $f_{m,n}=g_{m,n}$. This is because the offers are made at this level, so those who exit get $f_{m,n}$ in expectation. Also, those who did not get the offer, move to the state $(m-1,n-1)$ on the same diagonal, where the value functions are the same as in state (m,n).

Proof. Using the one-deviation property, assume that in all other states except (m,n) agents behave as mentioned above, and show that for each agent behavior in state (m,n) is optimal. If (m,n) lies above or below the main diagonal, the proof is straightforward. The interesting case is on the diagonal, for example if $(m,n)=(1,1)$. Then as in Proposition 31, define $f^{\infty}_{1,1}$ the expected payoff of the sellers if they all wait, and by $g^{\infty}_{1,1}$ the same notion for the buyers. One calculates

$$f^{\infty}_{1,1}=\frac{1}{2}(f_{1,2}+f_{2,1})=\frac{1}{2}(g_1+f_1)=\frac{1}{2},$$

since $f_1+g_1=1$ (by definition f and g were chosen symmetric with respect to $\frac{1}{2}$). Then the discounted value to time zero is $f^{0}_{1,1}=f^{\infty}_{1,1}-\frac{\varepsilon}{2}$ for the sellers, and $g^{0}_{1,1}=g^{\infty}_{1,1}+\frac{\varepsilon}{2}$ for the buyers. Since $f^{0}_{1,1}<g^{0}_{1,1}$, it follows that both the buyer and the seller prefer not to wait, and would make each other an offer in the interval $[\frac{1}{2}-\frac{\varepsilon}{2},\frac{1}{2}+\frac{\varepsilon}{2}]$. In particular $f_0=g_0=\frac{1}{2}$ belongs to this interval. □

Notice that other offers can be made in that interval, and it depends on agents beliefs. That suggests that the equilibrium is not unique. However, one can show that it is essentially unique. Recall that a rigid equilibrium has all boundary states partial rigid, i.e., in those states only the bottom traders use mixed strategies—which implies that the value functions are equal for the buyers and the sellers.

Proposition 13. *Any rigid equilibrium in the homogeneous case must collapse onto the coordinate axes.*

Proof. The boundary of the state region is by definition formed only with partial rigid states, where $f=g$. Now from a state (m,n) the system can only go to $(m,n+1)$ or $(m+1,n)$, so $2f_{m,n}+\varepsilon=f_{m+1,n}+f_{m,n+1}$, which means that f keeps increasing as it moves closer to the boundary. Similarly, g is decreasing towards the boundary. But at the boundary $f=g$, so in state (m,n) it must have been that $f<g$, which means that it is not

optimal for agents to wait in state (m, n). So there are no states in Ω except for the partial ones. But then the same argument can be used for a partial state, so it is not optimal to wait there either. □

8 Empirical Implications

8.1 Market Orders and the Spread

An interesting implication of the equilibrium in the general case is that a market sell order leads to an decrease in both the bid and the ask, but the decrease in the bid is larger. The intuition for this is given in Proposition 28: the departure of one limit buyer makes the buyers better off, and the sellers worse off (because of future possibility of trading with each other when the book becomes full). But that possibility is remote enough so that the decrease in the value function of the sellers is less than the decrease in the value function of the buyers (recall that the value function is minus the utility for the buyers). This intuition also appears in the analysis of the one-sided case, cf. equation (9).

This implication was noticed by Biais, Hillion and Spatt (1995), in their analysis of the order flow in the Paris Bourse. Their analysis goes as follows: *"The downward shift in the book has two components, the decrease in the bid, merely reflecting that the large sale consumed the liquidity offered at that quote, and the subsequent decrease in the ask, reflecting the reaction of the market participants to the large sale. The decrease in the bid could be a transient decrease in the liquidity on this side of the book, or a permanent information adjustment. In our one-lag analysis, we cannot differentiate the two hypotheses. In contrast to the behavior of the bid, the decrease in the ask is likely to reflect the information effect."*

I argue that the decrease in the ask need not reflect an information effect. Indeed, it can simply be regarded as an adjustment made by the limit sellers, who, after the bid decreased, realize that they now have to wait more to execute their orders, and lower their offers accordingly. It is however quite hard to distinguish between the waiting costs story and the information story.

One may still argue that if information acquisition is not homogeneous across agents, there is actually a way of distinguishing between the two stories. Namely, consider the information story. If a large market sell order occurs, and this conveys information, then it

is reasonable to expect more new agents than current ones to take advantage of the change. This is because it is presumably harder for current sellers in the book to change their beliefs about the value of the asset. Also, they may prefer not to cancel their limit orders, since this operation involves some costs (private and sometimes financial). So in the information story one should expect more new agents to arrive, and fewer limit order cancellations. This means that, if market sell orders are followed by a lot of cancellations on the ask side of the book, then that would reinforce more the waiting costs story.

And it turns out that this is true: cancellations on the ask side of the book are particularly frequent after market sell orders. This in fact is one of the strongest findings of Biais et al. besides the diagonal effect (which is just serial correlation of different types of orders). They suggest a possible explanation based on hidden orders in the book at the Paris Bourse, but as I explained above, the waiting costs story seems to be a likelier explanation.

8.2 The Distribution of the Bid-Ask Spread

Consider again the context of Section 4.3 (with patient sellers arriving faster than impatient buyers). The market is a Markov system with transition matrix P as in equation (23). To calculate the distribution of the bid-ask spread, one needs to know the stationary probability that the system is in state m. Denote this by x_m. Define

$$a = \frac{\omega}{1-\omega}, \quad \alpha = \frac{1-\omega}{\omega}, \quad \beta = \frac{1}{2\omega - 1}.$$

Consider the row vector X with entries x_m. From the theory of Markov matrices, one knows that $XP = X$. Solving for X, one gets $(1-\omega)x_{m+1} = \omega x_m$ for all m, hence $x_m = C(\frac{\omega}{1-\omega})^m = Ca^m$. The components x_m must sum to one, so $C = \frac{a-1}{a^{M+1}-1}$. Then

$$x_m = \frac{a^{m+1} - a^m}{a^{M+1} - 1} \approx \alpha^{M-m} - \alpha^{M-m+1}.$$

To calculate the spread $a_m - B$, use the formula $f_{m-1} - f_m = a^{M-m}\varepsilon(\beta + u) - \varepsilon\beta$. From this point forward assume that there is no randomization in state M (so $s = 0$ and

$u = \frac{1}{1-\omega}$). Then one gets the formula $f_{m-1} - f_m = (a^{M-m+1} - 1)\varepsilon\beta$, which implies

$$\begin{aligned} a_M - B &= \varepsilon\beta\,(a-1), \\ a_{M-1} - B &= \varepsilon\beta\,\big((a-1) + (a^2-1)\big), \\ a_{M-2} - B &= \varepsilon\beta\,\big((a-1) + (a^2-1) + (a^3-1)\big), \\ &\vdots \end{aligned}$$

Notice that these spreads appear with stationary probabilities approximately equal to $1 - \alpha, \alpha - \alpha^2, \alpha^2 - \alpha^3, \ldots$, respectively. It would be interesting to compare this distribution with the empirically observed one (especially at the tails of the distribution).

I give an example of calculating the first moment of the distribution, which may be of empirical interest. The exact formula for the bid-ask spread $\bar{S}$ is

$$\bar{S} = \frac{\varepsilon\beta}{(a-1)(a^{M+1}-1)}\left((M+1)a^{M+3} - (M+3)a^{M+2} + (M+3)a - (M+1)\right).$$

This can be very well approximated by

$$\bar{S} \approx \varepsilon\beta a M.$$

Now I give an asymptotic formula for M when ε is small. Recall from (14) that M is obtained by solving $A - B + \varepsilon\beta M = \frac{\varepsilon\beta a^{M+1}}{a-1}(1-\alpha^M)$. When ε is small, M is large, so a^M is much larger than M. Then one can drop M from the above formula, and also approximate $1 - \alpha^M$ with zero. One then obtains

$$M \approx \frac{1}{\log a}\log\frac{1}{\varepsilon}.$$

So the average spread $\bar{S} \approx \beta\frac{a}{\log a}\varepsilon\log\frac{1}{\varepsilon}$, which implies that asymptotically

$$\bar{S} \sim \varepsilon\log\frac{1}{\varepsilon}.$$

Compare this with Farmer et al. (2003), where $\bar{S} \sim \varepsilon^{3/4}$. (Notice that their ε, which they call "granularity parameter" is the same—up to a constant—as our ε, if one assumes that

the patience r of agents trading in different stocks is the same.)

Part II

Multi-Stage Games in Continuous Time

In this Part the strategic interaction of agents is modeled as a multi-stage game with observed actions,[25] but in continuous time. The framework I use borrows heavily from the theory of repeated games in continuous time, as developed by Bergin and MacLeod (1993). Since there are also stochastic moves by Nature and entry decisions by new players, I extend their framework by using an idea from Simon and Stinchcombe (1989).

Since this is a trading game, one may wonder why I consider simultaneous moves (organized in the "stage game" at each time t), and not have only successive moves by players. (Indeed, on organized exchanges there is a strict priority of both market and limit orders.) The reason is that it is technically easier to describe a game in continuous time where agents move simultaneously; otherwise, the submission order would have to be decided by a randomization procedure. Also, one may want successive actions to happen "instantly", which would force one to consider them all at the same t.

There is however a cost to be paid: extra care has to be taken about market orders. To see why, suppose at some time t a trader submits a market buy order, and simultaneously the bid and ask prices are determined by the limit orders placed by the other players. Then in equilibrium all sellers would compete to place a limit order at the top price A. This contradicts the intuition that the market order should be executed at the existing ask—which was set right before t—and not at the new ask set at t. One way to get around this problem is to demand that market orders have to be cleared at the bid or ask prices set right before t (of course, one will have to define precisely what "right before" means in continuous time). The other problem with market orders is that they reduce the number of players in the game. Since I want this event to happen at a definite time, I require that when a market order is submitted at t, the game between the remaining players takes place also at t. One way to get around this problem is by introducing "layered times" (see below), i.e.,

[25] See Fudenberg and Tirole (1991), chapter 4.

by allowing multiple games to played at a discrete set of times in $[0, \infty)$. This is reminiscent of auction theory, and the notion of "stopping the clock."

Now I discuss multi-stage games with observed actions in continuous time. I start with the case of pure strategies, and later in the section I deal with mixed strategies. Game theory in continuous time is not a straightforward extension of game theory in discrete time. There are a few conceptual difficulties, as pointed out by Simon and Stinchcombe (1989), or Bergin and MacLeod (1993). To understand why, suppose one tries to replicate a typical punishment strategy from discrete time repeated games:

> *Continue* to cooperate if the other player has not defected yet; if the other player defected at any point in the past, *immediately* defect and continue to defect forever.

The difficulty to make this strategy precise is two-fold. First, in continuous time there is no first time after t, which makes it difficult to "continue" a certain course. One way to get around this problem is to allow strategies to have inertia. But this creates a second problem, since the other players can take advantage of inertia.[26] One way to allow players to react "immediately" is to enlarge the concept of strategy to include sequences of faster and faster responses. The mathematical concept that allows us to do that is *completion* with respect to a metric (see below).[27]

A third problem that arises in continuous time is that there is no first time before t. This issue is important when one needs a description of the game right before t. For example, in a trading game, suppose a trader submits a market buy order at t. This market order is supposed to be very fast, and not give time for the existing traders to change their limit orders—otherwise, all sellers would try to change their limit order at the highest possible level A. One can then model this by allowing the market order to be placed "immediately before" t, and the game at t will be played with one less player, namely the trader whose limit order was cleared. Since there is no first time before t, it is not obvious at which price the market order is to be executed. The solution of this problem is to allow only strategies that "behave well" immediately before any time t. The technical concept, borrowed from

[26] One could force the players to all have the same inertia, but then this would be equivalent to forcing the game to take place in discrete time.

[27] Another way to define immediacy is by using infinitesimal numbers, which is the mathematical field of non-standard analysis.

Simon and Stinchcombe (1989), is of a strategy with a uniformly bounded number of jumps (to be defined below).

In order to define a game, one must define the spaces of actions, outcomes, and strategies. The definitions follow closely those of Bergin and MacLeod (1993). I extend their framework in several directions: (i) there is a well-defined description of the game right before any time t; (ii) I allow for entry decisions of new agents; and (iii) I account for the possibility of having more than one game played at the same time. I start with an infinitely countable set of players I. Since they arrive according to independent Poisson processes, with probability one at each point in time there are only finitely many traders. I want to include the case where at some times t the game is played more than once. I do this by taking the product of the time interval $[0, \infty)$ with the set of natural numbers $\mathbb{N}$, to indicate how many time a game has been played at some time t. Define

$$\mathcal{T} = [0, \infty) \times \mathbb{N}$$

the set of times at which players can move, counted with multiplicity. Notice that if $\leq$ is the lexicographic order, $(\mathcal{T}, \leq)$ is a totally ordered space. Denote the element $(0, 0) \in \mathcal{T}$ also by 0. Define intervals in $\mathcal{T}$ in the usual way: for example, if $T = (t, n) \in \mathcal{T}$ define $[0, T) = \{T' \in \mathcal{T} \mid 0 \leq T' < T\}$. When there is no danger of confusion, write t instead of (t, n). Also, define a measure on $[0, \infty)$ so that bounded measurable functions are integrable:

$$\mu(\mathrm{d}t) = \mathrm{e}^{-at}\mathrm{d}t.$$

In general, I want the action space for player i to be a compact complete metric space (X_i, d_i). Typically, X_i is a compact subset of $\mathbb{R}^n$ and d_i is the inherited metric. In the present case, the action space for player $i \in I$ can be defined as a subset of $\mathbb{R}^2$:

$$X_i = \big([B, A] \times \{0, 1\}\big) \cup \{\mathbf{out}\},$$

where **out** is some point in $\mathbb{R}^2$ which does not lie on $[B, A] \times \{0, 1\}$. An action $(x_i, 1) \in X_i$ is interpreted as a limit order at $x_i \in [B, A]$. An action $(x_i, 0) \in X_i$ is interpreted as a market sell (buy) order, in which case x_i is the current bid (ask) price, respectively. The

action $x_i = \mathbf{out}$ indicates that either (i) player i has not entered the game yet; or (ii) player i exited the game before. One could also allow agents to exit freely at time t. This will not happen in equilibrium if the utility from exiting is very small, so in order to simplify the description of the game I do not allow free exit. Define also projections on the first and second factor, $\pi_1 : X_i \to [B, A] \cup \{\mathbf{out}\}$ and $\pi_2 : X_i \to \{0, 1\} \cup \{\mathbf{out}\}$, in the obvious way.

I now define outcomes of the game. Let $\mathcal{B}_{X_i}$ and $\mathcal{B}_X$ be the Borel sets of X_i and $X = \prod_{i \in I} X_i$, respectively; and let $\mathcal{B}$ be the Borel sets of $[0, \infty)$. A function $\nu : [0, \infty) \to \mathbb{N}$ is said to have finite support if ν is zero everywhere except on a finite set $M_1 = \{t_1, \ldots, t_K\}$ (its support). One also associates the set $M = \{(t_1, n_1), \ldots, (t_K, n_K)\}$, where all $n_k = \nu(t_k) > 0$. Vice versa, for any such set M one can define a function $\nu_M : [0, \infty) \to \mathbb{N}$ with finite support by sending $t \in [0, \infty)$ to zero if $t \notin M_1$; and to n_k if $t = t_k \in M_1$.

Definition 6. *Let X be a space with measure. A function $f : \mathcal{T} \to X$ is called* ***layered*** *if there exists a function $\nu : [0, \infty) \to \mathbb{N}$ with finite support such that $\forall t \in [0, \infty)$ and $\forall n, n' > \nu(t)$ one has $f(t, n) = f(t, n')$. If $f : \mathcal{T} \to X$ is layered, associate a function $f^\nu : [0, \infty) \to X$ by $f^\nu(t) = f(t, \nu(t))$. I say that f is a layered measurable function if f^ν is measurable. An* ***outcome*** *for player i is a layered Borel measurable function $h_i : \mathcal{T} \to X_i$.*

So an outcome is like a regular measurable function $h_i : [0, \infty) \to X_i$, except that at a finite set $\{t_1, \ldots, t_K\}$ (the support of ν) it can take several values, up to the integer number $\nu(t_k)$. This corresponds to the idea that at some times t_k the game can be played more than once (in my case, if some agent places a market order).

I call the function ν the **layer** of f. Sometimes I also call the layer of f the associated set $M = \{(t_1, n_1), \ldots, (t_K, n_K)\}$, with $n_k = \nu(t_k)$. Also, if f_1 and f_2 are two layered functions with layers ν_1 and ν_2, one can take the combined layer of f_1 and f_2 to be $\nu = \max\{\nu_1, \nu_2\}$. This is useful for situations where one has to compare f_1 and f_2. Consider a layer ν. Then I define: $\mathcal{T}^\nu$, the set of layered times associated with ν; H_i, the space of outcomes for player i; and H_i^ν, the space of outcomes associated with ν:

$$\mathcal{T}^\nu = \{(t, n) \in \mathcal{T} \mid n \leq \nu(t)\},$$

$$H_i = \{h_i : \mathcal{T} \to X_i \mid h_i \text{ layered measurable}\},$$

$$H_i^\nu = \{h_i : \mathcal{T} \to X_i \mid h_i \text{ layered measurable with layer } \nu\}.$$

This is a metric space with the metric $D_i : H_i^\nu \times H_i^\nu \to \mathbb{R}_+$ given by

$$D_i(h_i, h_i') = \int_{[0,\infty)} d_i\big(h_i^\nu(t), h_i'^\nu(t)\big)\mu(\mathrm{d}t) + \sum_{k=1}^{K}\sum_{n=0}^{\nu(t_k)} d_i\big(h_i(t_k, n), h_i'(t_k, n)\big).$$

Rewrite this as

$$D_i(h_i, h_i') = \int_{T^\nu} d_i\big(h_i(T), h_i'(T)\big)\ \mu^\nu(\mathrm{d}T).$$

Since the space of measurable functions $f_i : [0,\infty) \to X_i$ is compact and complete, so is H_i^ν. Now, if $\nu \leq \nu'$, there is an inclusion $H_i^\nu \to H_i^{\nu'}$. Also, one knows that for every two layers ν_1 and ν_2 one can take their maximum $\nu = \max\{\nu_1, \nu_2\}$, which satisfies $\nu_1, \nu_2 \leq \nu$. This means that one can regard H_i as the limit of H_i^ν when ν becomes larger and larger. Because of this, H_i is a metric space, but it might not be either complete or compact.

I now define the space H of outcomes of the game. For this, let $H^\nu = \prod_{i\in I} H_i^\nu$ the product space with the metric $D = \prod_{i\in I} \frac{1}{2^i} D_i$. It is a standard exercise in measure theory to see that H^ν is compact and complete. As before, if $\nu \leq \nu'$, there is an inclusion $H^\nu \to H^{\nu'}$. I then define H as the union of all H^ν for all layers ν. This is still a metric space, but it might not be complete or compact. To justify this definition, consider an outcome $h \in H$. Since h belongs to a union of H^ν over all layers ν, there must exist a particular ν so that $h \in H^\nu$ (in which case, I say that ν is the layer of h). This corresponds to the fact that all agents are in the same game, played at the times described by the layer ν.

Also, if $Z \subset T^\nu$ is layered measurable, and $h_i, h_i' \in H_i^\nu$, define a metric relative to Z by $D_i(h_i, h_i', Z) = \int_Z d_i\big(h_i(T), h_i'(T)\big)\ \mu^\nu(\mathrm{d}T)$. Define also a metric D on H relative to Z in the same way it was done for the product metric above.

Now I define strategies. In discrete time, pure strategies map histories to actions, while mixed strategies map histories to probability densities over actions. For technical reasons it is easier to think of a history as an outcome of the game together with a time t at which history is taken. This way, one can define a strategy as a map from {outcomes $\times$ times} to {actions}. Formally, a strategy for agent i is a map

$$s_i : H \times \mathcal{T} \to X_i$$

which satisfies the following axioms

A1. The function s_i is layered measurable on $H \times \mathcal{T}$.

A2. For all $h, h' \in H$ and $T \in \mathcal{T}$ such that $D(h, h', [0, T)) = 0$, one has $s_i(h, T) = s_i(h', T)$.

The second axiom ensures that future does not affect current decisions. Rewrite

$$h \sim_T h' \iff D(h, h', [0, T)) = 0.$$

As it was discussed above, these two axioms alone do not ensure that strategies uniquely determine outcomes. For that, one needs some inertia condition. If $t \in [0, \infty)$ and ν is a layer, denote by $t^\nu = (t, \nu(t))$, and $t = (t, 0)$.

A3. The function s_i displays inertia, i.e., for any $h \in H^\nu$ and any $t \in [0, \infty)$, there exists $\varepsilon > 0$ and $x_i \in X_i$ such that

$$D_i\big(s_i(h'), x_i, [t^\nu, t + \varepsilon)\big) = 0$$

for every $h' \in H^\nu$ such that $h \sim_{t^\nu} h'$.

Denote by S_i the set of functions s_i on $H \times \mathcal{T}$ which satisfy A1, A2, A3. Denote by $S = \prod_{i \in I} S_i$. The next theorem shows that a strategy profile $s = (s_i)_i$, i.e., a set of strategies s_i for for each player $i \in I$, uniquely determine an outcome on every subgame. More precisely one has the following result:

Proposition 14. *Let $s \in S$. Then for every $h \in H$ and $T \in \mathcal{T}$, there exists a unique (continuation) outcome $\tilde{h} \in H$ so that $h \sim_T \tilde{h}$ and $D(s(\tilde{h}), \tilde{h}, [T, \infty)) = 0$.*

Proof. The proof is the same as in Bergin and MacLeod, but one has to make sure that one works in H^ν for some layer ν. □

Given $(h, T) \in H \times \mathcal{T}$ and $s \in S$, denote by $\sigma(s, h, t)$ the outcome which agrees with h on $[0, T)$ and is determined by the strategy s on $[T, \infty)$. Let $s_i, s_i' \in S_i$. I now define a metric on S_i:

$$\rho_i(s_i, s_i') = \sup_{H \times \mathcal{T} \times S_{-i}} D\big(\sigma((s_i, s_{-i}), h, T), \sigma((s_i', s_{-i}), h, T)\big).$$

One also has to introduce an axiom which ensures that for each t the outcome of the game right before t is well defined. One way of doing this is to restrict to strategies s_i that lead to locally constant outcomes with a uniformly bounded number of jumps.

A4. For the strategy s_i there exists M (depending only on s_i) such that for any strategies s_{-i} of the other players, the outcome $\sigma_i\big((s_i, s_{-i}), h, t\big)$ for player i is locally constant and has at most M jumps.

Redefine S_i to include on the strategies that satisfy A4. Now recall that at each $t \in [0, \infty)$ the strategies have inertia for some ε (depending on t). I want inertia to be infinitesimal, because I want to allow for immediate responses. This can be done by completing the space of strategies: Denote by S_i^* the completion of S_i with respect to the metric ρ_i, and by $S^* = \prod_{i \in I} S_i^*$. Completion is done so that the upper bound for the number of jumps is the same for all. More precisely, a point in S_i^* corresponds to a Cauchy sequence $(s_i^n)_n$ of strategies in S_i, and one demands that there exists M so that for each n, s_i^n jumps at most M times, regardless of the other players' strategies. The following result shows that to each strategy in S^* one can associate a unique outcome in H.

Proposition 15. *For every $s \in S^*$ and every $(h, T) \in H \times \mathcal{T}$, there exists a unique h^* such that $\sigma(s^n, h, T) \to h^*$ for any Cauchy sequence $(s^n)_n$ in S converging to s.*

Proof. If $(h, T) \in H \times \mathcal{T}$, there exists a layer ν so that $h \in H^\nu$ and $T \in \mathcal{T}^\nu$. The result then follows easily since H^ν is compact and complete. □

I have just showed that for $s \in S^*$ one can associate a unique outcome of the (whole) game, which I denote by $\sigma^*(s)$. Because completion is done using the same upper bound for the number of jumps, the following result is straightforward. The result allows one to talk about the outcome of a game right before some time t.

Proposition 16. *The outcome $\sigma^*(s)$ associated to a strategy $s \in S^*$ is left continuous.*

I am almost done in defining the game. The only thing that is left is to describe the payoff for some strategy $s \in S^*$ in a subgame defined by a history $(h, T) \in H \times \mathcal{T}$. Since strategies uniquely define outcomes in every subgame, as long as there exists some payoff $u_i(\sigma^*(s, h, T))$ for each agent i. Define now the equilibrium concept:

Definition 7. *A strategy profile $s \in S^*$ is an ε-Nash equilibrium (ε-NE) if for any $h \in H$*

$$u_i\big(\sigma^*(s,h,0)\big) \geq u_i\big(\sigma^*((s_i',s_{-i}),h,0)\big), \quad \forall i \in I, \quad \forall x_i' \in S_i^*.$$

A strategy profile $s \in S^*$ is an ε-subgame perfect Nash equilibrium (ε-SPE) if for any $(h,T) \in H \times \mathcal{T}$

$$u_i\big(\sigma^*(s,h,T)\big) \geq u_i\big(\sigma^*((s_i',s_{-i}),h,T)\big), \quad \forall i \in I, \quad \forall x_i' \in S_i^*.$$

For $\varepsilon = 0$ in the above inequalities one obtains the concepts of Nash equilibrium (NE) and subgame perfect Nash equilibrium (SPE). One has the following important result.

Proposition 17. *A strategy profile $s \in S^*$ is a subgame perfect equilibrium if and only if for any Cauchy sequence $(s^n)_n$ converging to s, there is a sequence $\varepsilon^n \to 0$ such that s^n is an ε^n-subgame perfect equilibrium.*

I now discuss mixed strategies, or rather, as they will be called, "pure-mixed." For simplicity of discussion I omit the presence of layers, so one takes $\mathcal{T} = [0,\infty)$. Consistent with this philosophy of locally constant outcomes and inertia strategies, I want to have mixed strategies randomly switch over a small interval. More formally, let X_i be the space of actions for player i, and $[0,\infty]$ the metric space with metric $d(x,y) = |\mathrm{e}^{-x} - \mathrm{e}^{-y}|$. Define a mixed strategy to be a layered measurable function

$$s_i : H \times \mathcal{T} \to X_i \times X_i \times [0,\infty],$$

where the first component of s_i is the initial action in X_i; the second component is the action to which s_i will switch in the interval of time right after t; the third component is the Poisson intensity of switching. I call this type of strategy **pure-mixed**, because randomness only comes from the time of switching, while the actions before or after switching are deterministically chosen. One could in principle also allow for randomization over these actions, in which case one has to replace X_i by $\Phi(X_i)$, the set of probability densities over X_i, i.e., the set of non-negative integrable functions on X_i with total integral equal to one. But then the analysis would become much more complicated, and these strategies are not

really necessary for the games considered in this thesis.

I say that the strategy s_i has inertia in a similar way as before, but one adds the requirement that after switching the value to which player i switched will be also held constant for a small period of time. Then one has to modify the description of outcomes, which are now stochastic processes and are built in a very similar way to Poisson processes. The space of strategies is also constructed by taking a completion, in the same way it was done for pure strategies.

To see how a pure-mixed strategy works, consider the case of Nature, which moves at each time t. The space of actions for Nature is the set 2^I of all subsets of I (in principle I allow Nature to add or remove any players from the game). Nature plays the following strategy: if $(h,t) \in H \times \mathcal{T}$ is the history at t, the first component of $s_N(h,t)$ is the set $J = I_{t-}$ of players right before t; the third component is the sum $\lambda = \lambda_{PS} + \lambda_{PB} + \lambda_{IB} + \lambda_{IS}$ of the arrival intensities of all four types of traders (patient sellers, etc.); the second component is a density over 2^I given by $\phi\big(J \cup \{PB\}\big) = \frac{\lambda_{PB}}{\lambda}$, $\phi\big(J \cup \{PS\}\big) = \frac{\lambda_{PS}}{\lambda}$, $\phi\big(J \cup \{IB\}\big) = \frac{\lambda_{IB}}{\lambda}$, $\phi\big(J \cup \{IS\}\big) = \frac{\lambda_{IS}}{\lambda}$, and $\phi(Z) = 0$ for any other set $Z \in 2^I$.

I now briefly discuss the notions of equilibrium employed in this thesis. The main one is that of a sub-game perfect equilibrium (SPE), which reflects the dynamic nature of the game. But, in order to formalize some other intuitions which are important for this game, I also consider several refinements. As discussed before, there can be many SPE of this game, for example the ones based on Nash threats (as in the theory of repeated games). However, if one focuses on "competitive" equilibria, i.e., equilibria where agents do not take into consideration what effect their action has on the other players' strategies, one sees that the the competitive equilibria are essentially unique (very close to each other).

In this context the notion of competitive equilibrium is very close to that of a Markov perfect equilibrium[28] (MPE), where the state variable is (m,n), with m the number of sellers and n the number of buyers active in the game at a particular time. A Markov or equilibrium is **stationary** if the value function only depends on the state (m,n) and not

[28]Let k_t be a variable which summarizes what one wants to know about a game at time t. Suppose k_t evolves according to a Markov process with transition function $q(k_{t+\mathrm{d}t}|k_t, a_t)$, which gives the probability that the next period's state is $k_{t+\mathrm{d}t}$ conditional on its being k_t at time t and on playing the actions a_t. By definition, if two histories lead to the same k_t, a Markov strategy maps them to the same action. A Markov perfect equilibrium is a perfect equilibrium in Markov strategies. Notice that the concept of MPE depends on the particular state variable chosen. See Fudenberg and Tirole (1991, ch. 13).

on time t.

Competitive equilibria are important in a trading game framework, because they give the main intuition for the equilibrium limit order book. One can think of a competitive equilibrium as a Bertrand-type competition among patient traders for the stream of market orders of the impatient traders. Since execution of limit orders is subject to the FIFO (First-In-First-Out) policy, it follows that traders have different times-to-execution for their orders, and that means that in order to compete (and get the same expected payoff), patient agents have to place limit orders at different levels in the book. I then say that an SPE is **competitive** if at each time t all agents have the same value function (possibly depending on t). I want to stress the fact that this definition is quite ad-hoc and is only valid in my context, where the symmetry of the payoffs is broken by the FIFO policy.

Part III

A CAPM with Price Impact (joint with A.W. Lo and J. Wang)

9 Introduction to Part III

The existence of large traders changes puts under question one typical assumption of asset pricing, namely that traders are price takers. The existence of price impact of transactions is a well-known fact in financial markets, and most large traders often consider the question of how to choose their trading strategies when taking into account their price impact. In this work, we investigate how the existence of price impact changes the consumption and investment behavior of such rational large agents.

We start by considering a market where there exist large traders who are assumed to have price impact, which is common knowledge. The setup of the model is very similar to a standard investment–consumption portfolio problem in discrete infinite time. We assume that all agents have constant absolute risk aversion, and asset returns are normally distributed. There is no information asymmetry in our model, so in order to give agents a

reason to trade, we assume that some agents receive shocks to wealth each period, shocks which are correlated to dividends. This is a reasonable assumption, since in reality agents' wealth may also have a non-financial component, which is nevertheless correlated with dividends.

The only nonstandard feature of our model is that some agents have a price impact of trading, while the rest do not. Price impact is modeled by a price function which is linear in demand. While the coefficients of demands in the price function are assumed exogenous, the constant term is determined in equilibrium. For tractability, we also assume that the agents who receive the wealth shock are the same as the ones with price impact.

We show that in equilibrium average prices display a liquidity premium, which depends on the price impact coefficient. An interesting phenomenon, in contrast with the intertemporal CAPM model, is that the holdings of the agents with price impact matter in equilibrium, even though all agents have absolute risk aversion (so wealth and asset holdings should not matter). In our model, the large agent holdings become a state variable, and prices in each period depend on the holdings of the large agents in the previous period. A subtle point of our analysis is that it is hard to disentangle the liquidity premium from the risk premium in our model. However, a Taylor series expansion can be made to do just that.

We show that the results are similar those in the transaction costs literature, where small transaction costs determine only small changes in equilibrium prices, but possibly large variations in equilibrium portfolio holdings. Indeed, we show that in the two-asset case, the equilibrium holdings in one asset might even be negative. The transaction costs literature assumes that, with each transaction, a trader incurs costs which are typically fixed or proportional to the size of the order. While realistic, this assumption leads to complicated technical problems, mainly given by the fact that either if a trader buys or sells an asset, the costs go in the same direction, so they depend on the absolute value of the size of the trade. Our approach has the advantage that the pricing function depends linearly on the size of the order, so it makes the solution of the model much easier, while preserving the intuition.

10 The model

The economy has a single good that can be used either for consumption or investment, and it can be regarded as numeraire. The model is a discrete time, infinite horizon model, which means that trading takes place at discrete times $t = 1, 2, 3, \ldots$. There is no final period, and investors are infinitely lived.

A. Investment opportunities

There is one riskless asset in the economy, with an infinitely elastic supply at a positive constant rate of return r. Denote by $R = 1 + r$ its gross rate of return.

There are also n risky assets (stocks) in this economy, each with fixed supply normalized to one. Each share of the risky asset $i = 1, \ldots, n$ pays a random dividend $D_{i,t}$ at time t. Denote by D_t the (column) n-vector $\begin{bmatrix} D_{1,t} & \ldots & D_{n,t} \end{bmatrix}^\top$. We assume that D_t is independently and identically distributed (i.i.d.), and we write

$$D_t = D + \varepsilon_{D,t}, \tag{49}$$

where D is a constant vector, and $\varepsilon_{D,t}$ is a zero-mean vector of "disturbances," whose exact distribution will be specified when we discuss the distributional assumptions. We will see later that different specifications for the dividend process (for example random walk, or $AR(1)$) would not lead to qualitatively different results.

B. Investors

There are two agents in this economy, one which we call C (the competitive agent), and one which we call N (the non-competitive agent). Both have constant absolute risk aversion (CARA), with absolute risk aversion coefficients α_C and α_N, respectively. Their time discount coefficients are, respectively, β_C and β_N, both numbers between 0 and 1. This means that the "felicity" function of agent $i = C, N$ is $u(c) = -\exp(-\alpha_i c)$, and that in each period t agent i maximizes expected intertemporal utility of consumption:

$$\mathsf{E}_t \left\{ -\sum_{s=0}^{\infty} \beta_i^s \exp(-\alpha_i c_{t+s}^i) \right\}. \tag{50}$$

We further assume that agent N has price impact in the risky assets (hence the term "non-competitive"), while agent C does not have any price impact. The exact mechanism of price impact will be described shortly. We also note that, since agent C is a price taker with CARA utility, the representative agent framework holds, so that having only one competitive agent in our model is general enough.

To give investors rational incentives to trade, we assume that agent N receives a shock to wealth F_t every period t. The shock arrives prior to trading at t, and is public knowledge. We assume the following functional form for the shock:

$$F_t = A_{t-1} \cdot \varepsilon_{F,t}, \tag{51}$$

where the "amplitude" A_t of the shock is an $AR(1)$ process given by

$$A_t = a_A A_{t-1} + \varepsilon_{A,t}, \text{ with } a_A \in [0, 1). \tag{52}$$

Here $\varepsilon_{F,t}$ and $\varepsilon_{A,t}$ are assumed to be i.i.d.

C. Equilibrium mechanism

Denote by X_t^N the n-vector of risky asset holdings of agent N after trading at t, and by X_t^C the total risky asset holdings of agent C. Denote by x_t^N the net demand of agent N at t, and similarly define x_t^C for agent C. To simplify notation, we will omit the superscript N. We have the formula $X_t^N = X_t = X_{t-1} + x_t$.

Denote by P_t the n-vector of (ex-dividend) prices of risky assets at which investors trade in period t. We assume that P_t is of the form

$$P_t = V_t + \lambda \cdot x_t. \tag{53}$$

where V_t is an n-vector determined in equilibrium, and λ is a constant $n \times n$ matrix determined exogenously to the model. We call λ the price impact coefficient. A consequence of this functional form for P_t is that in the dynamic optimization problem for agent N, the variable x_t becomes a control variable, so this is the sense in which we can say that N has price impact.

One may object to the fact that P_t is not determined by some market mechanism, and that instead we impose a functional form on it. However, one can make this mechanism result endogenously as follows: Suppose agent N submits a market order for x_t shares, and agent C submits the whole demand function, which we assume is linear in price: $x_t^C(p) = a - bp$ (true if C has CARA utility and returns are normal). Then a Walrasian auctioneer will set the price p so that demand equals supply, i.e., $a - bp + x_t = 0$. This implies $p = \frac{a}{b} + \frac{1}{b}x_t$, which is of the desired form. However, we prefer to be agnostic as to the source of the price impact, and start with our specification (53) for the pricing mechanism, with λ exogenous. Then we can concentrate on how the equilibrium prices change as we let λ vary, which is the main point of this approach.

Another justification of the above price specification can be given by looking at a one-period model with the same type of price impact. Assume there are only two agents in this model, which have both CARA utility with the same coefficient of absolute risk aversion γ. As above, we assume that N has price impact, and C does not. There exists a riskless asset with gross rate of return R, and one risky asset (a stock) with normally distributed final-period payoff, $q \sim N(\bar{q}, \sigma^2)$. The price mechanism is the same as above. Denote by S the supply of the risky asset, by h the initial stock holdings of agent N (before trading), by x the net stock demand, by X the final stock holdings, and by P the trading price. Then one can calculate

$$x = \frac{\frac{S}{2} - h}{1 + \lambda \frac{R}{2\gamma\sigma^2}}, \tag{54}$$

$$P = \frac{1}{R}(\bar{q} - \frac{\gamma S}{2}\sigma_q^2) - \frac{\lambda}{2}\frac{\frac{S}{2} - h}{1 + \lambda \frac{R}{2\gamma\sigma_q^2}}. \tag{55}$$

When $\lambda = 0$, one finds that $x = \frac{S}{2} - h$, and $X = \frac{S}{2}$, which represents the usual optimal risk sharing. However, as λ becomes very large, x becomes close to zero, indicating that there will be very little trade due to adverse large price impact. This suggests that the above price specification does induce agents to behave as we would expect them to. Notice also that the equilibrium price P converges to a finite value when λ goes to infinity, which is a phenomenon that also appears in the infinite period model.

Define Q_t to be the excess share return of the risky assets. This is the n-vector given

by the formula

$$Q_t = P_t + D_t - RP_{t-1}. \tag{56}$$

Each component of Q_t represents the return on one share of the corresponding stock, minus the cost of financing at the riskless rate. Note that Q_t is not the same thing as the excess *rate* of return. To obtain the latter, one should divide the share return Q_t by the price P_{t-1}. Since after we make our distributional assumptions it may be the case that P_{t-1} is zero, we prefer not to divide by price, and work with the share return Q_t instead.

D. DISTRIBUTIONAL ASSUMPTIONS

We denote by ε_t the $(n+2)$-vector of random shocks, $\varepsilon_t = \begin{bmatrix} \varepsilon_{D,t} & \varepsilon_{A,t} & \varepsilon_{F,t} \end{bmatrix}^\top$. We assume that ε_t is i.i.d. multivariate normal, with zero mean and covariance matrix

$$\Sigma = \mathsf{E}\{\varepsilon_t \varepsilon_t^\top\} = \begin{bmatrix} \sigma_{DD} & o & \sigma_{DF} \\ o^\top & \sigma_{AA} & 0 \\ \sigma_{DF}^\top & 0 & \sigma_{FF} \end{bmatrix}. \tag{57}$$

The notation $\sigma_{XY} = \mathsf{E}\{\varepsilon_{X,t}\varepsilon_{Y,t}^\top\}$ represents the covariance matrix of $\varepsilon_{X,t}$ with $\varepsilon_{X,t}$. The covariance matrix σ_{DD} is $n \times n$, while σ_{DF} is $n \times 1$, and σ_{AA}, σ_{FF} are scalars. Note that we assumed that $\varepsilon_{D,t}$ and $\varepsilon_{A,t}$ are uncorrelated (o is the zero vector), while $\varepsilon_{D,t}$ and $\varepsilon_{F,t}$ are assumed to have covariance matrix σ_{DF}.

The reason we want $\varepsilon_{D,t}$ and $\varepsilon_{F,t}$ to be correlated is that otherwise the shocks F_t would be uncorrelated to the dividends, and so they would have no impact on portfolio decisions. That would imply that there is no trade after the first period.

E. ORDER OF EVENTS

The order of events that happen at time t is as follows: First, all agents receive their dividends per share D_t, and $\varepsilon_{D,t}$ is revealed. Then agent N receives the wealth shock F_t, and $\varepsilon_{F,t}$ and $\varepsilon_{A,t}$ are revealed. At this point agents make their consumption–investment decisions, and consume c_t^i, $i = N, C$. Then both agents submit their demands x_t^i, when their stock holdings are X_{t-1}^i. At the same time with agents' submission of x_t^i, the market maker (who is a Walrasian auctioneer) announces V_t. The actual trading price P_t is then determined by the formula $P_t = V_t + \lambda \cdot x_t$. After trading takes place, the stock holdings

become X_t^i. We recall that there is no asymmetric information in this model, so all news is public knowledge.

11 Equilibrium

We now solve for the equilibrium of the model in Section 3. Since the problem is not very different from a standard Merton intertemporal portfolio choice problem, we are only going to dwell on the points where our model differs.

The general idea is to start with some parametric guesses for the price process and the value functions for agents N and C. Then, from the optimization problems for the two agents plus the market clearing equation in each risky asset, we should determine the values of those parameters and show that our guesses were correct.

A. The equilibrium price

We begin by making a guess about the price process. We assume that it must be linear in a set of variables that influence both agents' demands. Now, agent C has CARA utility and is a price taker, which means that once the price process is determined, C's optimal demand will depend on whatever the price depends on, and no other variables are needed to determine C's choice. Therefore, the price process is guided by the variables that determine the optimal demand of agent N. We are going to discuss this issue in more depth below, when we analyze the optimization problem of agent N. For now, we just give some intuition in order to justify our guess for the price process.

First, since all agents have CARA utility, N's wealth process is not going to affect the price process, although it is a state variable for N's optimization problem (that's one of the reasons, besides tractability, why we chose CARA utility in the first place). Since N's wealth shock at $t+1$ has amplitude A_t and is correlated with dividends, N's hedging demand at t depends on A_t, so we assume that A_t is one of the state variables that influence the price process. Also, since N has price impact, we should expect one more state variable: X_{t-1}, N's stock holdings before trading at t. The precise reason why X_{t-1} is a state variable will be explained when we discuss the wealth process of agent N, as part of finding N's optimal portfolio.

In what follows, by state variable we mean one of the variables that determine the price process and agents' optimal demands and value functions at each time t. We saw in the above discussion that A_t and X_{t-1} are state variables. To simplify notation, we also introduce the constant 1 as a state variable. So the $(n+2)$-vector of state variables is

$$Z_t = \begin{bmatrix} 1 & A_t & X_{t-1}^\top \end{bmatrix}^\top. \tag{58}$$

In the next theorem we show that there exists an equilibrium vector of stock prices P_t which depends linearly on Z_t. The state variables also evolve in a linear fashion.

Theorem 18. *The economy described in the previous section has a stationary*[29] *equilibrium vector of stock prices given by*

$$P_t = p_1 + p_A A_t + p_X X_{t-1} = pZ_t, \tag{59}$$

where p_1 and p_A are constant $n \times 1$ vectors, and p_X is a constant $n \times n$ matrix. The state variables Z_t evolve according to the equation

$$Z_{t+1} = a_Z Z_t + b_Z \varepsilon_{t+1}, \tag{60}$$

where a_Z and b_Z are constant $(n+2) \times (n+2)$ matrices.

Proof. See Appendix B. □

This result describes prices in terms of the state variables, and shows how the state variables evolve in time. Notice that the only part of Z_{t+1} that we did not know how it evolves is X_t, the vector of N's stock holdings after trading at t. We can directly give a recursive formula for X_t, or equivalently we can give a formula for the net stock trading vector, $x_t = X_t - X_{t-1}$:

$$x_t = q_1 + q_A A_t + q_X X_{t-1} = qZ_t. \tag{61}$$

Now that we know both P_t and x_t, we can also determine the equilibrium value for V_t,

[29] There can be also non-stationary equilibria, as we shall see in Appendix B. These equilibria are also linear in the state variable Z_t and have finite mean, but their variance blows up when t becomes large.

which is the vector present in the price mechanism equation (53):

$$V_t = P_t - \lambda x_t = (p - \lambda q)^\top Z_t. \tag{62}$$

B. Optimal portfolios

We now discuss the optimal demands and value functions of agents N and C. As in any dynamic optimization problem, we need to find for each agent a set of state variables, a set of control (choice) variables, and, since there is uncertainty in the problem, a set of random disturbances. These variables are related through the state equation, which shows how the system evolves in time.

In a standard Merton portfolio problem, the state variables include those variables that determine the investment opportunities set (Z_t in our case), together with the investor's wealth process W_t. By wealth one means total wealth, i.e., the total dollar value of holdings in both the stocks and the riskless asset. Since in a standard model there is no price impact of trading, total wealth is a well-defined concept (trading does not change the total value of holdings). Denote by c_t the consumption of an investor before trading at t, and by X_t the investor's vector of stock holdings right after trading at t. Clearly c_t and X_t are control variables. Then total wealth evolves according to the following formula

$$W_{t+1} = (W_t - c_t)R + (X_t)^\top (P_{t+1} + D_{t+1} - RP_t).$$

This is part of the state equation (the other part describes the evolution of the other state variables, such as Z_t). Notice that instead one could choose the net demand $x_t = X_t - X_{t-1}$ as a control variable. But in that case, we get an extra state variable, X_{t-1}, which unnecessarily complicates the problem.

Coming back to our model, notice that the above wealth equation suffers from two problems in the case of agent N. One is that, by equation (53), P_t now explicitly depends on x_t, which is a control variable. This means that we cannot avoid X_{t-1} becoming a state variable anymore. But, more seriously, it does not make sense to consider W_t as a state variable, since it depends on x_t (or X_t), which is a control variable contemporaneous to W_t. This reason induces us to look for a state variable which does not depends on x_t.

I. Optimization for N

An appropriate state variable for the optimization problem of agent N is w_t, the before-trading cash wealth, calculated immediately before trading at time t. The cash wealth incorporates the shock $F_t = A_{t-1}\varepsilon_{F,t}$, hence it satisfies

$$w_{t+1} = (w_t - c_t - x_t^\top P_t)R + (X_{t-1} + x_t)^\top D_{t+1} + A_t\varepsilon_{F,t+1}. \tag{63}$$

This formula shows that w_t, A_t and X_{t-1} are indeed state variables, x_t and c_t are control variables, and $\varepsilon_{D,t+1}$, $\varepsilon_{F,t+1}$ are random disturbances.

To complete the state equation for agent N, we also need to describe how the other state variables evolve, i.e., to give a recursive formula for Z_{t+1}. But $Z_{t+1} = \begin{bmatrix} 1 & A_{t+1} & X_t^\top \end{bmatrix}^\top$, so we only have to give a recursive formula for X_t. One may think that we already did this in (61), which implies that in equilibrium $X_t = q_1 + q_A A_t + (q_X + I)X_{t-1}$. However, this is not how agent N regards the evolution of X_t, since N has control over it. More precisely, X_t depends on X_{t-1} via the control variable x_t: $X_t = X_{t-1} + x_t$. This gives a different equation for N, of the form

$$Z_{t+1} = a_Z^N Z_t + b_Z^N \varepsilon_{t+1} + c_Z^N x_t. \tag{64}$$

(The exact formulas for the coefficients will be given in Appendix B.) Together, equations (63) and (64) form the state equation for agent N.

For the next result, we make the guess that the value function of agent N at t is log-linear in cash wealth and log-quadratic in Z_t. Then using the Bellman principle of optimality[30], one calculates the optimum consumption and stock holdings.

Theorem 19. *Let $w_t^N = w_t$ be the cash wealth of agent N before trading at time t, c_t^N his consumption, $x_t^N = x_t$ the net stock demand, and J_t^N his value function. His optimization problem is*

$$J_t^N(w_t, Z_t) = \max_{c^N, x^N} E_t\left\{ -\sum_{s=0}^{\infty} \beta_N^s \exp\left(-\alpha_N c_{t+s}^N\right) \right\},$$

[30]Suppose that we want to solve the dynamic optimization problem $\mathsf{E}_0 \max_{(x_t)_t} \sum_{t=0}^{\infty} f(k_t, x_t)$, such that the state equation $k_{t+1} = g(k_t, x_t, u_t)$ holds for all t. Here k_t is the state variable, x_t is the control variable, and u_t is a random disturbance. The value function is defined as $J_t(k_t) = \max_{x_s} \mathsf{E}_t \sum_{s=t}^{\infty} f(k_s, x_s)$, subject to the state equation. Then Bellman's principle says that $J_t(k_t) = \max_{x_t} \Big(f(k_t, x_t) + \mathsf{E}_t J_{t+1}(k_{t+1}) \Big)$.

subject to the state equations (63) and (64), and a transversality condition. It has the following solution

$$J_t^N = -\beta_N^t \exp\left(-\gamma_N w_t - \tfrac{1}{2} Z_t^\top v^N Z_t\right), \tag{65}$$

$$c_t^N = \tfrac{r}{R} w_t - \tfrac{1}{\alpha_N R} \ln(r\beta_N \delta_N) + \tfrac{1}{\alpha_N R}\, \tfrac{1}{2} Z_t^\top g^N Z_t, \tag{66}$$

$$x_t = h^N Z_t, \tag{67}$$

where $\gamma_N = \alpha_N \frac{r}{R}$ *and* δ_N *are scalars,* v^N *and* g^N *are constant symmetric* $(n+2) \times (n+2)$ *matrices, and* h^N *is a constant* $n \times (n+2)$ *matrix.*

Proof. See Appendix B. □

II. Optimization for C

Agent C behaves as in the standard Merton portfolio problem, so as we discussed at the beginning of this section, we can use total wealth W_t^C and Z_t as state variables, and the vector of stock holdings X_t^C and consumption c_t^C as control variables. The recursive formula for the total wealth of C is

$$W_{t+1}^C = \left(W_t^C - c_t^C\right) R + \left(X_t^C\right)^\top (P_{t+1} + D_{t+1} - RP_t). \tag{68}$$

This is the first part of the state equation. The second part describes the evolution of Z_t from agent C's perspective. Since C has no control over Z_t, the recursive formula for Z_{t+1} is the same as (60), which is the equilibrium one.

Theorem 20. *Let* W_t^C *be the total wealth of agent* C, c_t^C *his consumption,* X_t^C *the total stock holdings after trading at* t, *and* J_t^C *his value function. His optimization problem is*

$$J_t^C(W_t, Z_t) = \max_{c^C, X^C} E_t \left\{ -\sum_{s=0}^{\infty} \beta_C^s \exp\left(-\alpha_C c_{t+s}^C\right) \right\},$$

subject to the state equations (68) and (60), and a transversality condition. It has the

following solution

$$
\begin{aligned}
J_t^C &= -\beta_C^t \exp\bigl(-\gamma_C W_t^C - \tfrac{1}{2} Z_t^\top v^C Z_t\bigr), && (69)\\
c_t^C &= \tfrac{r}{R} W_t^C - \tfrac{1}{\alpha_C R} \ln(r\beta_C\delta_C) + \tfrac{1}{\alpha_C R}\, \tfrac{1}{2} Z_t^\top g^C Z_t, && (70)\\
X_t^C &= h^C Z_t, && (71)
\end{aligned}
$$

where $\gamma_C = \alpha_C \frac{r}{R}$ *and* δ_C *are scalars,* v^C *and* g^C *are constant symmetric* $(n+2)\times(n+2)$ *matrices, and* h^C *is a constant* $n \times (n+2)$ *matrix.*

Proof. See Appendix B. □

C. FINAL EQUATIONS

We now briefly show how to solve explicitly for the equilibrium described in Theorems 18–20 (the full details will be given in Appendix B). We have to choose a set of unknowns which describes the solution, and indicate how to find equations that determine these unknowns. A natural choice for the set of unknowns should be based on the guesses we made, so we choose p, q, v^N and v^C. The $(n+2)$-vector p comes from equation (59): $P_t = pZ_t$; the $(n+2)$-vector q comes from equation (61): $x_t = qZ_t$; and the $(n+2)\times(n+2)$ matrices v^N and v^C come from the value functions J^N and J^C in equations (65) and (69).

The first set of equations comes from the formula for the optimal net stock demand of agent N, which is $x_t = h^N Z_t$. The second set of equations comes from market clearing, which is $X_t^N + X_t^C = u$, where u is the n-vector of ones (we normalized stock supply to one). Then we have two sets of equations for v^N and v^C coming from the Bellman principle. We finally get the following system of equations:

$$
\begin{aligned}
q &= h^N,\\
f_u &= \bigl(h^N + f_X\bigr) + h^C, && (72)\\
v^N &= \tfrac{1}{R} g^N - 2\ln\Bigl(\tfrac{R}{r}\bigl(r\beta_N\delta_N\bigr)^{1/R}\Bigr) E^{11},\\
v^C &= \tfrac{1}{R} g^C - 2\ln\Bigl(\tfrac{R}{r}\bigl(r\beta_C\delta_C\bigr)^{1/R}\Bigr) E^{11},
\end{aligned}
$$

where f_u and f_X are constant $n\times(n+2)$ matrices (defined in Appendix B); and E^{11} is the $(n+2)\times(n+2)$ matrix which has the top left entry equal to one and all others equal to zero.

In order solve for the equilibrium, one has to find a solution to this nonlinear system of equations. How many unknowns do we have? In the last two sets of equations, v^i, g^i and E^{11} are symmetric matrices, so only above-diagonal elements matter. In conclusion, we have in total $2(n+2)+2(n+2)(n+3)/2=(n+2)(n+5)$ independent equations, with the same number of unknowns.

We discuss the solution to this system of equations in the next section, as we also interpret the results.

12 Analysis of Results

Before we indicate how to solve the system of equations in the previous section, let us discuss the stationarity of the equilibrium processes P_t and X_t. As we shall see below, the solution to our nonlinear system is unique as long as we require that the resulting processes P_t and X_t be stationary.

A. Stationarity

We would like to be able to describe P_t and X_t as $ARMA$ processes, and see if they are stationary. Let for simplicity $n=1$ (the results are true in general). Recall that the solution we are interested in has the form

$$P_t = p_1 + p_A A_t + p_X X_{t-1}, \tag{73}$$

$$X_t = q_1 + q_A A_t + q'_X X_{t-1}, \tag{74}$$

where $q'_X = q_X + 1$. Denote by $\eta_t = \varepsilon_{A,t}$. Recall that A_t is an $AR(1)$ process given by $A_t = \gamma A_{t-1} + \eta_t$. We want to describe X_t and P_t as $ARMA$ processes, so we demean them, which means that we ignore the free terms p_1 and q_1. Start with the equation $X_t - q'_X X_{t-1} = q_A A_t$. Subtract from this the same equation, but lagged by one and multiplied by γ. We get

$$X_t - (q'_X + \gamma)X_{t-1} + \gamma q'_X X_{t-2} = q_A \eta_t, \tag{75}$$

which means that X_t is an $AR(2)$ process. A standard time series result implies that X_t is

stationary if and only if the polynomial $\gamma(L) = 1 - (q_X + \gamma)L + \gamma q'_X h^2$ has roots outside the unit circle. In our case, since $\gamma \in (0,1)$, it follows that X_t is stationary if and only if $q'_X \in (-1,1)$.

To describe P_t, denote by $A_t = X_t - \gamma X_{t-1}$, and by $B_t = P_t - \gamma P_{t-1}$. From equation (75), we get that $A_t - q'_X A_{t-1} = q_A \eta_t$, so A_t is an $AR(1)$ process. But $P_t = p_A A_t + p_X X_{t-1}$, so if we subtract the same equation lagged by one and multiplied by γ, we get $B_t = p_A \eta_t + p_X A_{t-1}$. We can now calculate $B_t - q'_X B_{t-1} = p_A \eta_t + (p_X q_A - p_A q'_X)\eta_{t-1}$. But since $B_t = P_t - \gamma P_{t-1}$, we can finally write

$$P_t - (q'_X + \gamma)P_{t-1} + \gamma q'_X P_{t-2} = q_A \eta_t + (p_X q_A - p_A q'_X)\eta_{t-1}, \tag{76}$$

which shows that P_t is an $ARMA(2,1)$ process.

We have the same conditions for the stationarity of P_t as for X_t (the right hand side of an $ARMA$ process does not matter for stationarity). Also, it is clear from the description of P_t above that the excess share return $Q_t = P_t + D_t - (1+r)P_{t-1}$ is an $ARMA(2,2)$ process.

B. Solving the system

In the previous section we saw that a linear solution to our model of price impact can be found via Theorems 18–20 by solving the system (72) of nonlinear equations. We denote the unknowns by $\mathbf{x}$, which is a vector composed by the vectors p, q, and the super-triangular parts of the symmetric matrices v^N and v^C. The system then can be written under the form $\mathbf{F}(\mathbf{x}) = 0$, where the vector $\mathbf{F}(\mathbf{x})$ is composed by $\mathbf{F}_{opt^N}$, $\mathbf{F}_{mkt}$, and the super-triangular part of $\mathbf{F}_{val^N}$ and $\mathbf{F}_{val^C}$, which are defined by

$$\begin{aligned}
\mathbf{F}_{opt^N} &= q - h^N, \\
\mathbf{F}_{mkt} &= f_u - \left(h^N + f_X\right) - h^C, \\
\mathbf{F}_{val^N} &= v^N - \tfrac{1}{R} g^N + 2\ln\left(\tfrac{R}{r}\left(r\beta_N \delta_N\right)^{1/R}\right) E^{11}, \\
\mathbf{F}_{val^C} &= v^C - \tfrac{1}{R} g^C + 2\ln\left(\tfrac{R}{r}\left(r\beta_C \delta_C\right)^{1/R}\right) E^{11}.
\end{aligned} \tag{77}$$

The first set of equations $\mathbf{F}_{opt^N} = 0$ comes from the optimization problem of agent N; the

second set of equations $\mathbf{F}_{mkt} = 0$ comes from market clearing; the third and fourth set of equations come from the Bellman equation for the value functions of agents N and C;

Let us look at the case when there is only one risky asset, i.e., $n = 1$. Then the system $\mathbf{F}(\mathbf{x}) = 0$ is an 18×18 nonlinear system of equations. Let $\mathbf{x}$ be its solution, which we write as

$$\begin{aligned} \mathbf{x} \;=\; \Big[\; & p_1 \quad p_A \quad p_X \quad q_1 \quad q_A \quad q_X \quad \dots \\ & v_{11}^N \quad v_{12}^N \quad v_{13}^N \quad v_{22}^N \quad v_{23}^N \quad v_{33}^N \quad v_{11}^C \quad v_{12}^C \quad v_{13}^C \quad v_{22}^C \quad v_{23}^C \quad v_{33}^C \;\Big]^\top . \end{aligned} \tag{78}$$

The solution $\mathbf{x}$ clearly depends on the parameters of the system. These parameters are: r, the risk-free rate (and $R = 1 + r$); α_N and α_C, the coefficients of absolute risk aversion of both agents (we are also going to use the parameters $\gamma_i = \frac{\alpha_i r}{R}$); β_N and β_C, the time-discount coefficients; D, the mean value of dividends; σ_{DD}, the variance of dividends; σ_{AA}, the variance of $\varepsilon_{A,t}$ (the innovation to A_t, the amplitude of the wealth shock); σ_{FF}, the variance of $\varepsilon_{F,t}$; σ_{DF}, the covariance of $\varepsilon_{D,t}$ with $\varepsilon_{F,t}$; and γ, the autocorrelation of A_t. We change notation:

$$\phi = a_A, \qquad \sigma_{11} = \sigma_{DD}, \qquad \sigma_{22} = \sigma_{AA}, \qquad \sigma_{33} = \sigma_{FF}, \qquad \sigma_{13} = \sigma_{DF}. \tag{79}$$

We also give more notation, by defining some new parameters γ_0 and ω instead of γ_N and γ_C:

$$\gamma_0 = \frac{\gamma_N \gamma_C}{\gamma_N + \gamma_C}, \qquad \omega = \frac{\gamma_C}{\gamma_N + \gamma_C}. \tag{80}$$

Even though we are in the simplest case, $n = 1$, the system is quite complicated, so it is doubtful that one can find a solution in closed form, except for some special values of the parameters. However, it can easily be solved numerically, using the standard Newton–Raphson method (see Press et al (1992), chapter 9). The only observation here is that in order to speed up calculations it is important to use an analytic formula for the Jacobian of the system. Otherwise the Jacobian has to be approximated numerically, and this considerably slows down the Newton–Raphson algorithm.

As it is usual when solving systems of nonlinear equations, one needs to start the algorithm by finding an 18-dimensional vector $\mathbf{x}_0$ which is reasonably close to the solution.

The best way of doing that is by setting some parameters equal to zero and then trying to solve the system analytically. The two things that complicate the system are the wealth shock and the price impact, so in order to find $\mathbf{x}_0$ we set $\sigma_{22} = 0$ and $\lambda = 0$. Although the solution $\mathbf{x}$ depends on all parameters, from now on we only indicate the dependence on λ and σ_{22}; we also define the solution for when λ and σ_{22} are zero:

$$\mathbf{x} = \mathbf{x}(\lambda, \sigma_{22}) \qquad \mathbf{x}_0 = \mathbf{x}(0, 0).$$

The condition $\sigma_{22} = 0$ means that $\varepsilon_{A,t} = 0$, which implies that $A_t = \gamma A_{t-1} + \varepsilon_{A,t}$ is also zero, so there is no wealth shock. The condition $\lambda = 0$ means that both agents are price takers, so we are in the usual Merton framework of the optimal consumption and investment problem. We take this as our benchmark, and analyze it first.

I. Benchmark case: no shocks, no price impact

This problem is standard, except that we have introduced two extra state variables, A_t and X_{t-1}, so we should in principle calculate their coefficients as well. But if we only look at the values of P_t and X_t, it turns out that we can calculate them without worrying about A_t and X_{t-1}. This is shown in the following result.

Proposition 21. *Consider the model where there is one risky asset, no wealth shock ($\sigma_{22} = 0$) and no price impact ($\lambda = 0$). Denote by X_{-1} the initial holdings of agent N. Then the model has a unique equilibrium, given by*

$$\begin{aligned} X_t &= \omega \quad \textit{for all} \quad t \geq 0, \\ P_t &= \frac{D - \gamma_0 \sigma_{11}}{r} \quad \textit{for all} \quad t \geq 0. \end{aligned}$$

This implies that trading only takes place at $t = 0$, after which agents reach their optimal risk-sharing holdings.

Proof. See Appendix B. □

Since we want to use this benchmark case to find a starting point $\mathbf{x}_0$ for the algorithm in the general case, we also have to express the solution as $P_t = p_1 + p_A A_t + p_X X_{t-1}$ and

$X_t = q_1 + q_A A_t + (q_X + 1)X_{t-1}$. Moreover, we also want to calculate v^N and v^C. Then $\mathbf{x}_0$ is the vector formed with $p_1, p_A, p_X, q_1, q_A, q_X$ and the super-diagonal elements of v^N and v^C.

Proposition 22. *A solution $\mathbf{x}_0$ to the nonlinear system (72) in the degenerate case when $\sigma_{22} = 0$ and $\lambda = 0$ is given by:*

$$p_1 = \frac{D - \gamma_0\sigma_{11}}{r} \qquad p_A = -\frac{\sigma_{13}\gamma_0}{R-\phi} \qquad p_X = 0 \qquad q_1 = \omega \qquad q_A = -\frac{\sigma_{13}(1-\omega)}{\sigma_{11}} \qquad q_X = -1$$

$$v_{11}^N = \frac{\gamma_0^2\sigma_{11} - 2R\ln(\frac{R}{r}) - 2\ln(r\beta_N)}{r} \qquad v_{12}^N = -\frac{\sigma_{13}\gamma_0^2(1-\omega)}{(R-\phi)\omega} \qquad v_{13}^N = \frac{(D-\gamma_0\sigma_{11})\gamma_0}{r\omega}$$

$$v_{22}^N = -\frac{\left(\sigma_{11}\sigma_{33} - \sigma_{13}^2(1-\omega)^2\right)\gamma_0^2}{\sigma_{11}(R-\phi^2)\omega^2} \qquad v_{23}^N = -\frac{\sigma_{13}\gamma_0^2}{(R-\phi)\omega} \qquad v_{33}^N = 0$$

$$v_{11}^C = \frac{\gamma_0^2\sigma_{11} - 2R\ln(\frac{R}{r}) - 2\ln(r\beta_C)}{r} \qquad v_{12}^C = \frac{\sigma_{13}\gamma_0^2}{R-\phi} \qquad v_{13}^C = 0$$

$$v_{22}^C = \frac{\sigma_{13}^2\gamma_0^2}{\sigma_{11}(R-\phi^2)} \qquad v_{23}^C = 0 \qquad v_{33}^C = 0$$

This is the unique solution if we make the extra requirement that q_X be negative, which corresponds to the processes X_t and P_t having a stationary representation.

Proof. See Appendix B. □

II. General case. Approximation of the solution

Now that we know the solution $\mathbf{x}_0$ when $\lambda = 0$ and $\sigma_{22} = 0$, it makes sense to look at the Taylor expansion of the solution $\mathbf{x}$ for general λ and σ_{22} around zero. Denote the long-term averages of X_t and P_t by

$$\bar{X} = \mathsf{E}X_t \quad \text{and} \quad \bar{P} = \mathsf{E}P_t. \tag{81}$$

Recall that $P_t = p_1 + p_A A_t + p_X X_{t-1}$ and $X_t = q_1 + q_A A_t + (q_X + 1)X_{t-1}$. Using these formulas, we get

$$\bar{X} = -\frac{q_1}{q_X} \quad \text{and} \quad \bar{P} = p_1 + p_X\bar{X}. \tag{82}$$

We now give the Taylor approximation for $\bar{P}$ around $\lambda = 0$ and $\sigma_{22} = 0$. Here the index 1 indicates the parameter λ, and 2 indicates the parameter σ_{22}.

Proposition 23. *For small values of λ and σ_{22}, the Taylor expansion of $\bar{P}$ (the long-term average of the equilibrium price) is*

$$\begin{aligned}\bar{P}(\lambda,\sigma_{22}) &= a_\emptyset + a_1\cdot\lambda + a_2\cdot\sigma_{22} + a_{11}\cdot\lambda^2 + a_{12}\cdot\lambda\sigma_{22} + a_{22}\cdot\sigma_{22}^2 \\ &\quad + a_{111}\cdot\lambda^3 + a_{112}\cdot\lambda^2\sigma_{22} + \cdots + a_{1111}\cdot\lambda^4 + \cdots, \end{aligned} \tag{83}$$

where there exist factors t_{22} and t_{112} depending only on R, ω, γ such that

$$a_\emptyset = \frac{D-\gamma_0\sigma_{11}}{r} \qquad a_1 = 0 \qquad a_2 = -\frac{\sigma_{13}^2\gamma_0^3}{r(R-\phi)^2} \tag{84}$$

$$a_{11} = 0 \qquad a_{12} = \frac{\sigma_{13}^2\gamma_0^2(1-\omega)\omega\big(2r+(1-\omega)(1-\phi)\big)}{r\sigma_{11}(R-\phi)^2} \qquad a_{22} = \frac{\sigma_{13}^2\gamma_0^3(1-\omega)^2}{r\sigma_{11}(R-\phi)^4(R-\phi^2)}\cdot t_{22}$$

$$a_{111} = 0 \qquad a_{112} = \frac{\sigma_{13}^2\gamma_0(1-\omega)^2\omega}{r\sigma_{11}^2(R-\phi)^2}\cdot t_{112} \qquad \cdots \qquad a_{1111} = 0 \qquad \cdots.$$

The factor t_{22} is always negative. The factor t_{112} is negative if $\omega \le 0.872$, while if $\omega > 0.872$, it is still negative for most values of $R > 1$ and $\phi \in (0,1)$.

For small λ and σ_{22}, the Taylor expansion of $\bar{X}$ (the long-term average of the equilibrium holdings of agent N) is

$$\begin{aligned}\bar{X}(\lambda,\sigma_{22}) &= b_\emptyset + b_1\cdot\lambda + b_2\cdot\sigma_{22} + b_{11}\cdot\lambda^2 + b_{12}\cdot\lambda\sigma_{22} + b_{22}\cdot\sigma_{22}^2 \\ &\quad + b_{111}\cdot\lambda^3 + b_{112}\cdot\lambda^2\sigma_{22} + \cdots + b_{1111}\cdot\lambda^4 + \cdots, \end{aligned} \tag{85}$$

where

$$b_\emptyset = 1 \qquad b_J = \frac{r(1-\omega)}{\gamma_0\sigma_{11}}\cdot a_J \quad \text{if length of index } J \text{ is at least one.} \tag{86}$$

Proof. See Appendix B. □

Notice that the Taylor expansion of $\bar{X}$ indicates a certain relationship between $\bar{P}$ and $\bar{X}$, hence a relationship among p_1, p_X, q_1, q_X.

Corollary 24. *If $\bar{P}$ and $\bar{X}$ are the average equilibrium price and N stock holdings, respectively, then*

$$\bar{P} - \frac{D-\gamma_0\sigma_{11}}{r} = k(\bar{X}-1), \qquad \text{where} \qquad k = \frac{\gamma_0\sigma_{11}}{r(1-\omega)}. \tag{87}$$

The constant k is independent of the the price impact parameter (λ) and the wealth shock

parameters (σ_{22}, σ_{33}, σ_{13}, γ).

More importantly, by looking at the Taylor expansion for p_1, p_X, q_1, q_X, one gets the following result:

Proposition 25. *Denote by X_{-1} the initial stock holdings of agent N. Then in the degenerate case when σ_{22}=0 (no wealth shock) but $\lambda > 0$ (agent N has price impact), we have the following formulas*

$$P_t - \frac{D - \gamma_0 \sigma_{11}}{r} = b(\lambda) a(\lambda)^t (X_{-1} - \omega), \tag{88}$$

$$X_t - \omega = a(\lambda)^{t+1} (X_{-1} - \omega), \tag{89}$$

where $a(\lambda)$ and $b(\lambda)$ are functions of λ such that $a(\lambda) \in (0,1)$. In particular, it follows that $\bar{P} = (D - \gamma_0 \sigma_{11})/r$ and $\bar{X} = \omega$.

Proof. See Appendix B. □

A Proofs of Results in Part I

Here is a proof of Proposition 2.

Proof. One needs to show that for each pair $A > B$ and each $\varepsilon > 0$ there exists a unique maximal chain. First, let us solve in general the difference equation $2f_m + \varepsilon = f_{m-1} + f_{m+1}$, with initial condition $f_0 = A$. This is a quadratic function

$$f_m = A - bm + \frac{\varepsilon}{2} m^2,$$

with b some arbitrary real constant that has to be determined from the terminal condition, $f_M = B$. Now solve $f_M = B$, i.e.,

$$A - Mu\varepsilon - M(M-1)\frac{\varepsilon}{2} = B$$

The positive solution is (if $R = \frac{2(A-B)}{\varepsilon}$)

$$M_u = \frac{1}{2} - u + \sqrt{(\frac{1}{2} - u)^2 + R}.$$

Notice that $M_0 = \frac{1}{2} + \sqrt{\frac{1}{4} + R}$ and $M_1 = -\frac{1}{2} + \sqrt{\frac{1}{4} + R}$, so $M_0 - M_1 = 1$. Therefore in the interval $[M_1, M_0)$ there is a unique integer M which corresponds to a number $u \in (0, 1]$. This yields M and u, and now $b = \varepsilon(M - \frac{1}{2} + u)$. □

I now prove Theorem 9. Assume that there is a rigid competitive stationary Markov equilibrium of the game. Then one proceeds to study necessary conditions for this equilibrium to exist.

Definition 8. *A state is called*

a) full *if agents leave the state only if some new agent arrives;*

b) partial *if an existing agent voluntarily exits after a positive expected waiting time;*

c) fleeting *if agents only stay in this state an infinitesimal time.*

A state is called regular *if it is either full or partial, so that agents will wait in it at least a positive amount of time.*

I now investigate the state region Ω, which is the collection of all regular states.

Proposition 26. *There exist a finite number of states in Ω (with probability one).*

Proof. Since agents lose utility proportionally to expected waiting time, there is a maximum expected time they will wait. Because of Poisson arrivals of the other agents, with probability one only a finite number of agents will arrive during this period (which is fixed in expectation). Now one has only to put this together with the fact that strategies of players in this game have all a uniformly bounded number of jumps (see Part II). □

Definition 9. *In any state S denote by f_S the value function of the sellers and by g_S the value function of the buyers.*

I already assumed the equilibrium is rigid, which means that at the boundary γ some agents behave using mixed strategies, in such a way that buyers and sellers have the same value function. Then one has the following result:

Proposition 27. *In a regular state $f_S \geq g_S$. In a partial or fleeting state $f_S = g_S$.*

Proof. To prove the first statement suppose in a regular state $f_S < g_S$. Then there exists $u > 0$ such that $f_S < g_S - u$. So one of the sellers is better off offering at $g_S - u$, which one of the buyers is better off accepting. For the second result, one knows that by definition in all partial states S one has $f_S = g_S$. Now each fleeting state S corresponds along the main diagonal to a partial state. Now use Proposition 31 to show that $f_S = g_S$ as well (it is the only possible equilibrium). □

Now I prove an important result, which allows one to determine the shape of the state region. It says that when a new seller arrives, every seller is worse off. Also, every buyer is better off, but by less than every seller is worse off.

Proposition 28. *Suppose the system is in a regular state S and a new seller arrives, so the system moves to regular state S'. Then $f_{S'} < f_S$ and $g_{S'} > g_S$. Moreover,*

$$f_S - f_{S'} \geq g_S - g_{S'}.$$

Proof. The difficult inequality is the last one. If S' is a partial state, then $f_{S'} = g_{S'}$ and the proof is done. So suppose S' is a full state. The statement must be hardest to prove for one buyer, so I show how to prove it in that case. Even in the most optimistic scenario

$$g_{S'} \geq B + \frac{\varepsilon}{2}.$$

Also, one can assume that no extra buyers are arriving (one can prove the statement for each number of buyers). Then

$$g_S \leq B + \varepsilon,$$

because the buyer just has to wait for an impatient seller. One gets

$$g_S - g_{S'} \leq \frac{\varepsilon}{2}.$$

Now for f: All the top agents in S' have to wait at least until an impatient buyer or patient (that becomes impatient) arrives into S. So one has

$$f_S - f_{S'} \geq \frac{\varepsilon}{2},$$

from which one deduces that $f_S - f_{S'} \geq g_S - g_{S'}$. □

Corollary 29. *If a state $S = (m, n)$ is regular and $m > 0$, $n > 0$, then $(m-1, n)$ and $(m, n-1)$ are also regular.*

Proof. Follows from the previous result. □

Let Ω be a region in the positive quadrant of the plane which is symmetric with respect to the main diagonal. Suppose Ω has the property that if (m, n) are in Ω, then so are $(m-1, n)$ and $(m, n-1)$. Let R be the largest integer such that $(0, R)$ is in Ω. Define a function $\phi : \{0, 1, \ldots, R\} \to \mathbb{N}$ in the following way: if $k = 0, 1, \ldots, R$ look at the intersection of the diagonal from the point $(0, R-k)$ with the boundary of Ω. This is a point of the form $(0+j, R-k+j)$ for some $j \in \mathbb{N}$. Then define $\phi(k) = j$. Call ϕ the associated function, or the shape function of Ω. The following result is not hard to prove.

Lemma 30. *Let Ω be a symmetric integer region in the positive quadrant, with the property that if (m, n) are in Ω, then so are $(m-1, n)$ and $(m, n-1)$. Then the shape function ϕ is always increasing by either 0 or 1. Conversely, any such function leads to a region Ω with the properties mentioned above.*

I now analyze more carefully what happens in each state (m, n). For this, take the one-sided story, and assume there are m sellers which compete for a bid of h. One can then prove the following important result, which is the game of attrition with Poisson arrivals.

Proposition 31. *Suppose m sellers lose utility in a way proportional to expected waiting time and coefficient r. At random time T which represents the first arrival in a Poisson process with intensity λ, an event happens and the game ends (this event can be the arrival of a new agent). Then, if all sellers wait until T, assume that each gets a payoff of f^∞. Also, at each time there exists a buyer who posts a bid for h. Assume that if a seller accepts h until T, he gets h and all other sellers get f^-. Denote by $f^0 = f^\infty - r/\lambda$. Then one has the following list of possible sub-game perfect equilibria:*

- *If $h > \max\{f^0, f^-\}$, then every seller immediately accepts h (and only one randomly gets it).*
- *If $h < \min\{f^0, f^-\}$, then no seller accepts h, and everybody waits until T.*

- *If $h \in [f^-, f^0]$ and $f^- \leq f^0$, there are two SPE: either everybody waits until T (this is the Pareto optimal equilibrium); or, if they believe the others will try to get h, they are all better off by doing the same: so each seller places a market order for h.*

- *$h \in [f^0, f^-]$ and $f^0 \leq f^-$, then this is a typical game of attrition. It has two equilibria: one, where some agent always accepts h, and the other never accept h; and the other where all agents accept h according to some Poisson process with intensity μ (μ is such that each agent is indifferent between accepting h now and waiting for the other $m-1$ sellers to do that).*

Now, the previous result does not assume anything about sellers placing limit orders. The next result is a simple extension of this game of attrition, where sellers are allowed to place limit orders. Clearly, the ask price is important now, because that might influence the payoff f^∞ at T. I show that one gets one more equilibrium.

Corollary 32. *In the setup of the previous proposition, suppose $\lambda = \lambda_1 + \lambda_2$, where λ_1 is the Poisson intensity of the arrival of an impatient buyer (event 1), and λ_2 corresponds to any other event that does not depend on the value of the ask price (event 2). Now suppose that at T, if event 1 happens, the bottom seller (at the ask) gets the ask price, while all the other get f_1^∞. If event 2 happens, assume that all sellers get f_2^∞. Then, besides the equilibria above there is one more equilibrium, where the top sellers wait and the bottom seller randomly accepts h with Poisson(μ), where μ is in such a way that everybody's value function is h.*

This last equilibrium indicates exactly how agents behave in partial rigid states.

Proposition 33. *Consider the state space Ω of a competitive stationary Markov equilibrium. Then, along each line parallel to the main diagonal there are initial full states, then at most one partial state, and then fleeting states. If the equilibrium is also rigid, then there is exactly one partial state.*

Proof. Follows from Proposition 28 and Corollary 32. □

Now I prove Theorem 11.

Proof. I show that $\triangle f = 1$. Start with equation $4f_{m,n} + \varepsilon = f_{m-1,n} + f_{m+1,n} + f_{m,n-1} + f_{m,n+1}$, and divide throughout by $\varepsilon = \delta^2$. Then one gets

$$\frac{f_{m-1,n} - 2f_{m,n} + f_{m+1,n}}{\delta^2} + \frac{f_{m,n-1} - 2f_{m,n} + f_{m,n+1}}{\delta^2} = 1.$$

But this is the finite difference approximation of the PDE

$$\left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\right)(m\delta, n\delta) = 1,$$

which is exactly $\triangle f(x, y) = 1$.

Now equation $3f_{m,0} + \varepsilon = f_{m-1,0} + f_{m+1,0} + f_{m,1}$ becomes after division by δ:

$$\frac{f_{m-1,0} - 2f_{m,0} + f_{m+1,0}}{\delta^2} \cdot \delta + \frac{f_{m,1} - f_{m,0}}{\delta} = \delta.$$

After passing to the limit when δ goes to zero, one gets

$$\frac{\partial f}{\partial y}(x, 0) = 0.$$

If one picks a point on γ of type 1, one has $(4 + s_{m,n})f_{m,n} + \varepsilon = 2f_{m-1,n} + 2f_{m,n-1} + s_{m,n}f_{m-1,n-1}$, which after division by δ becomes

$$2\frac{f_{m,n} - f_{m-1,n}}{\delta} + 2\frac{f_{m,n} - f_{m-1,n}}{\delta} + s_{m,n}\frac{f_{m,n} - f_{m-1,n-1}}{\delta} = -\delta.$$

After passing to the limit when δ goes to zero, one gets

$$\frac{\partial f}{\partial x}(x, y) + \frac{\partial f}{\partial y}(x, y) = 0.$$

For a point on γ of type 2, one has $(4 + s_{m,n})f_{m,n} + \varepsilon = f_{m-1,n} + 2f_{m,n-1} + f_{m,n+1} + s_{m,n}f_{m-1,n-1}$, which becomes

$$\frac{f_{m,n} - f_{m-1,n}}{\delta} + \frac{f_{m,n} - f_{m-1,n}}{\delta} + \frac{f_{m,n-1} - 2f_{m,n} + f_{m,n+1}}{\delta^2} \cdot \delta + s_{m,n}\frac{f_{m,n} - f_{m-1,n-1}}{\delta} = -\delta.$$

In the limit one gets the same condition $\frac{\partial f}{\partial x}(x, y) + \frac{\partial f}{\partial y}(x, y) = 0$.

Finally, the condition $f = g$ on γ is obvious. □

B Proofs of Results in Part III

Throughout the Appendix we use the following conventions and notations: Unless otherwise stated, vector notations represent vectors in column form, i.e., $n \times 1$ matrices. We denote the transpose of a matrix A by $A^\top$. Denote by $e_i^{(n)}$ the column n-vector with unit entry in position i and zero entries elsewhere; by $u^{(n)}$ the column n-vector with all entries equal to one; and by $o^{(n)}$ the zero n-vector. Also, denote by $O^{(m,n)}$ the $m \times n$ matrix with zero entries; by $O^{(n)}$ the zero $n \times n$ matrix; and by $I^{(n)}$ the identity $n \times n$ matrix. When there is no danger of confusion, we are going to omit the superscript n.

The general idea for proving Theorems 18–20 is the following: We first assume that the guesses made in these theorems are true. If we fix the parameter values for these guesses, the system is determined, and we only have to show that agents' behavior described in Theorems 19 and 20 is indeed optimal, and that our guesses for the value functions were correct. Moreover, we have to show that the price process postulated in Theorem 18 clears the market. These verifications lead to a set of equations in our initially fixed values of the parameters. If a solution can be found, then we are done.

With this strategy in mind, assume that the equilibrium price is given by $P_t = pZ_t$; that the equilibrium net stock demand of N is given by $x_t = qZ_t$, which determines the equilibrium evolution of Z_t, as in the state equation (60); and that the log-value functions for agents $i = N, C$ are quadratic in Z_t with Hessian matrix v^i. Then start with some fixed values for p, q, v^N and v^C.[31]

We now solve the optimization problems in Theorems 19 and 20. By using some more notation, we can tackle both optimization problems for agents N and C at the same time. Denote by $\tilde{W}_t^{\,N} = w_t$, the cash wealth of agent N before trading at t, and by $\tilde{W}_t^{\,C} = W_t^C$, the total wealth of agent C at t. Denote by $\tilde{X}_t^N = x_t$, the net stock demand of N at t, and by $\tilde{X}_t^C = X_t^C$, the total stock holdings of C after trading at t. This sets uniform notation for both agents: if $i = N, C$, then $\tilde{W}_t^{\,i}$ and Z_t are the state variables for agent i, while c_t^i and $\tilde{X}_t^i$ are the control variables. For example, equations (65) and (69), which express our

[31] Technically, we should also start with a value for γ_i, the coefficient of wealth in the log-value function, but it turns out that γ_i is equal to $\frac{\alpha_i r}{R}$, so we might as well assume it is known from the beginning.

guesses for the value functions of the two agents, give the following formula for the value function of agent i, $i = N, C$,

$$J_t^i = J_t^i(\tilde{W}_t^{\,i}, Z_t) = -\beta_i^t \exp\bigl(-\gamma_i \tilde{W}_t^{\,i} - \tfrac{1}{2} Z_t^\top v^i Z_t\bigr). \tag{90}$$

As we pointed out in Section 11, the state vector Z_t evolves differently from the viewpoint of the two agents. From the perspective of agent N, the state variables evolve according to the following equations: $1 = 1$, $A_{t+1} = a_Z A_t + \varepsilon_{A,t+1}$, $X_t = X_{t-1} + x_t$. For agent C, however, X_t evolves differently: $X_t = q_1 + q_A A_t + (q_X + I) X_{t-1}$. We can put these equations together and get the following state equation for agent $i = N, C$:

$$Z_{t+1} = a_Z^i Z_t + b_Z^i \varepsilon_{t+1} + c_Z^i \tilde{X}_t^i, \tag{91}$$

where

$$a_Z^N = \begin{bmatrix} 1 & 0 & o^\top \\ 0 & a_A & o^\top \\ o & o & I \end{bmatrix}, \qquad b_Z^N = \begin{bmatrix} o^\top & 0 & 0 \\ o^\top & 1 & 0 \\ O & o & o \end{bmatrix}, \qquad c_Z^N = \begin{bmatrix} o^\top \\ o^\top \\ I \end{bmatrix} \tag{92}$$

and

$$a_Z^C = \begin{bmatrix} 1 & 0 & o^\top \\ 0 & a_A & o^\top \\ q_1 & q_A & q_X + I \end{bmatrix}, \qquad b_Z^C = \begin{bmatrix} o^\top & 0 & 0 \\ o^\top & 1 & 0 \\ O & o & o \end{bmatrix}, \qquad c_Z^C = \begin{bmatrix} o^\top \\ o^\top \\ O \end{bmatrix}. \tag{93}$$

We also give some formulas that will be useful later on:

$$\begin{aligned} v_{aa}^i &= (a_Z^i)^\top v^i a_Z^i, \quad v_{bb}^i = (b_Z^i)^\top v^i b_Z^i, \quad v_{cc}^i = (c_Z^i)^\top v^i c_Z^i, \\ v_{ab}^i &= (a_Z^i)^\top v^i b_Z^i, \quad v_{ca}^i = (c_Z^i)^\top v^i a_Z^i, \quad v_{cb}^i = (c_Z^i)^\top v^i b_Z^i. \end{aligned} \tag{94}$$

Let us now look at the wealth processes of the two agents. Recall from equation (63) that the before-trading cash wealth of agent N evolves by $w_{t+1} = (w_t - c_t - x_t^\top P_t)R + (X_{t-1} + x_t)^\top D_{t+1} + F_{t+1}$. Using equation (53), we rewrite this as $w_{t+1} = (w_t - c_t)R - x_t^\top (V_t + \lambda x_t)R + (X_{t-1} + x_t)^\top (D + \varepsilon_{D,t+1}) + A_t \varepsilon_{F,t+1}$. We saw in equation (62) that $V_t = (p - \lambda q)^\top Z_t$. So we can then rewrite the wealth equation for N as follows: $w_{t+1} = (w_t - c_t)R - \frac{1}{2}(\tilde{X}_t^N)^\top m_{XX}^N \tilde{X}_t^N + (\tilde{X}_t^N)^\top m_{XZ}^N Z_t + (\tilde{X}_t^N)^\top m_{X\varepsilon}^N \varepsilon_{t+1} + \frac{1}{2} Z_t^\top m_{ZZ}^N Z_t + Z_t^\top m_{Z\varepsilon}^N \varepsilon_{t+1}$,

where the matrices m^N_{XX}, m^N_{XZ}, $m^N_{X\varepsilon}$, m^N_{ZZ}, $m^N_{Z\varepsilon}$ will be given below.

The wealth process of agent C is given in equation (68): $W^C_{t+1} = (W^C_t - c^C_t)R + \big(X^C_t\big)^\top (P_{t+1} + D_{t+1} - RP_t)$. Rewrite this as $W^C_{t+1} = (W^C_t - c^C_t)R - \big(X^C_t\big)^\top \big(p(Z_{t+1} - RZ_t) + D_{t+1}\big)$. Using equation (60), we can rewrite the wealth equation as follows: $W^C_{t+1} = (W^C_t - c^C_t)R - \frac{1}{2}(\tilde{X}^C_t)^\top m^C_{XX}\tilde{X}^C_t + (\tilde{X}^C_t)^\top m^C_{XZ}Z_t + (\tilde{X}^C_t)^\top m^C_{X\varepsilon}\varepsilon_{t+1} + \frac{1}{2}Z_t^\top m^C_{ZZ}Z_t + Z_t^\top m^C_{Z\varepsilon}\varepsilon_{t+1}$, where the matrices m^C_{XX}, m^C_{XZ}, $m^C_{X\varepsilon}$, m^C_{ZZ}, $m^C_{Z\varepsilon}$ will be given below.

We can put together the two wealth equations for agent $i = N, C$:

$$\begin{aligned} \tilde{W}^i_{t+1} &= (\tilde{W}_t^{\,i} - c^i_t)R + \tfrac{1}{2}(\tilde{X}^i_t)^\top m^i_{XX}\tilde{X}^i_t + (\tilde{X}^i_t)^\top m^i_{XZ}Z_t \\ &\quad + (\tilde{X}^i_t)^\top m^i_{X\varepsilon}\varepsilon_{t+1} + \tfrac{1}{2}Z_t^\top m^i_{ZZ}Z_t + Z_t^\top m^i_{Z\varepsilon}\varepsilon_{t+1}, \end{aligned} \tag{95}$$

where

$$\begin{gathered} m^N_{XX} = -2R\lambda, \qquad m^N_{XZ} = \begin{bmatrix} D & o & O \end{bmatrix} - R(p - \lambda q), \qquad m^N_{X\varepsilon} = \begin{bmatrix} I & o & o \end{bmatrix}, \\ m^N_{ZZ} = \begin{bmatrix} 0 & 0 & D^\top \\ 0 & 0 & o^\top \\ D & o & O \end{bmatrix}, \qquad m^N_{Z\varepsilon} = \begin{bmatrix} o^\top & 0 & 0 \\ o^\top & 0 & 1 \\ I & o & o \end{bmatrix}, \end{gathered} \tag{96}$$

and

$$\begin{gathered} m^C_{XX} = O, \qquad m^C_{XZ} = \begin{bmatrix} D & o & O \end{bmatrix} - Rp + pa^C_Z, \qquad m^C_{X\varepsilon} = \begin{bmatrix} I & p_A & o \end{bmatrix}, \\ m^C_{ZZ} = O^{(n+2)}, \qquad m^C_{Z\varepsilon} = O^{(n+2)}. \end{gathered} \tag{97}$$

We now solve the optimization problem of agent $i = N, C$, by applying the Bellman principle of optimality. If we denote by $J^i_t = J^i_t(\tilde{W}_t^{\,i}, Z_t)$ the value function of agent i, the Bellman principle states that

$$J^i_t = \max_{c^i_t, \tilde{X}^i_t} \Big(-\beta_i^t \exp(-\alpha_i c^i_t) + \mathsf{E}_t J^i_{t+1} \Big). \tag{98}$$

We need to calculate $\mathsf{E}_t J^i_{t+1}$. In order to do this, we use equations (95) and (91), which describe the evolution of the state variables $\tilde{W}_t^{\,i}$ and Z_t from the viewpoint of agent i. We

can write

$$J^i_{t+1} = -\beta_i^{t+1} \exp\bigl(-\gamma_i \tilde{W}^i_{t+1} - \tfrac{1}{2} Z^\top_{t+1} v^i Z_{t+1}\bigr) = -\beta_i^{t+1} \exp\bigl(-a - b^\top \varepsilon - \tfrac{1}{2}\varepsilon^\top A \varepsilon\bigr), \qquad (99)$$

where

$$\begin{aligned}
\varepsilon &= \varepsilon_{t+1}, \\
a &= \gamma_i (\tilde{W}_t{}^i - c^i_t) R + \tfrac{\gamma_i}{2} (\tilde{X}^i_t)^\top m^i_{XX} \tilde{X}^i_t + \gamma_i (\tilde{X}^i_t)^\top m^i_{XZ} Z_t + \tfrac{\gamma_i}{2} Z_t^\top m^i_{ZZ} Z_t \\
&\quad + \tfrac{1}{2} (\tilde{X}^i_t)^\top v^i_{cc} \tilde{X}^i_t + (\tilde{X}^i_t)^\top v^i_{ca} Z_t + \tfrac{1}{2} Z_t^\top v^i_{aa} Z_t, \\
b &= \bigl(\gamma_i m^i_{X\varepsilon} + v^i_{cb}\bigr)^\top \tilde{X}^i_t + \bigl(\gamma_i m^i_{Z\varepsilon} + v^i_{ab}\bigr)^\top Z_t, \\
A &= v^i_{bb}.
\end{aligned}$$

Now we apply the following standard lemma in multivariate normal calculus.

Lemma 34. *Let ε be a multivariate normal random variable, with zero mean and covariance matrix Σ. Let b be a constant vector, and A a constant symmetric semi-positive definite matrix. Define $\Omega = \bigl(A + \Sigma^{-1}\bigr)^{-1}$. Then*

$$E \exp\left(-b^\top \varepsilon - \tfrac{1}{2}\varepsilon^\top A \varepsilon\right) = \left(\frac{|\Omega|}{|\Sigma|}\right)^{1/2} \exp\left(\tfrac{1}{2} b^\top \Omega b\right), \qquad (100)$$

where $|M|$ denotes the determinant of the square matrix M.

Define $\Omega^i = \bigl(v^i_{bb} + \Sigma^{-1}\bigr)^{-1}$, and $\delta_i = \left(\frac{|\Omega^i|}{|\Sigma|}\right)^{1/2}$. We can calculate

$$\Omega^i = \begin{bmatrix} \sigma_{DD} & o & \sigma_{DF} \\ o^\top & \frac{\sigma_{AA}}{1+\sigma_{AA} v^i_{AA}} & 0 \\ \sigma^\top_{DF} & 0 & \sigma_{FF} \end{bmatrix} \qquad \text{and} \qquad \delta_i = \frac{1}{\bigl(1 + \sigma_{AA} v^i_{AA}\bigr)^{1/2}}, \qquad (101)$$

where v^i_{AA} is the $(n+1) \times (n+1)$ entry of the matrix v^i. We finally get

$$\mathsf{E}_t J^i_{t+1} = -\delta_i \beta_i^{t+1} \exp\Bigl(-\gamma_i R(\tilde{W}_t{}^i - c^i_t) - \tfrac{1}{2}(\tilde{X}^i_t)^\top u^i_{XX} \tilde{X}^i_t - (\tilde{X}^i_t)^\top u^i_{XZ} Z_t - \tfrac{1}{2} Z_t^\top u^i_{ZZ} Z_t\Bigr), \qquad (102)$$

where the matrices u^i_{XX}, u^i_{XZ} and u^i_{ZZ} are given by

$$\begin{aligned} u^i_{XX} &= \gamma_i m^i_{XX} + v^i_{cc} - \left(\gamma_i m^i_{X\varepsilon} + v^i_{cb}\right) \Omega^i \left(\gamma_i m^i_{X\varepsilon} + v^i_{cb}\right)^\top, \\ u^i_{XZ} &= \gamma_i m^i_{XZ} + v^i_{ca} - \left(\gamma_i m^i_{X\varepsilon} + v^i_{cb}\right) \Omega^i \left(\gamma_i m^i_{Z\varepsilon} + v^i_{ab}\right)^\top, \\ u^i_{ZZ} &= \gamma_i m^i_{ZZ} + v^i_{aa} - \left(\gamma_i m^i_{Z\varepsilon} + v^i_{ab}\right) \Omega^i \left(\gamma_i m^i_{Z\varepsilon} + v^i_{ab}\right)^\top. \end{aligned} \tag{103}$$

Define

$$g^i = u^i_{ZZ} - (u^i_{XZ})^\top (u^i_{XX})^{-1} u^i_{XZ} \quad \text{and} \quad h^i = -(u^i_{XX})^{-1} u^i_{XZ}. \tag{104}$$

Define the quadratic function Φ^i by

$$\Phi^i(\tilde{X}_t, Z_t) = \tfrac{1}{2} \tilde{X}_t^\top u^i_{XX} \tilde{X}_t + \tilde{X}_t^\top u^i_{XZ} Z_t + \tfrac{1}{2} Z_t^\top u^i_{ZZ} Z_t. \tag{105}$$

Now rewrite the Bellman equation (98) as follows

$$\begin{aligned} &-\beta_i^t \exp\left(-\gamma_i \tilde{W}_t^{\,i} - \tfrac{1}{2} Z_t^\top v^i Z_t\right) = \\ &\quad \max_{c_t^i, \tilde{X}_t^i} \left\{ -\beta_i^t \exp(-\alpha_i c_t^i) - \delta_i \beta_i^{t+1} \exp\left(-\gamma_i R (\tilde{W}_t^{\,i} - c_t^i) - \Phi^i(\tilde{X}_t^i, Z_T)\right) \right\}. \end{aligned} \tag{106}$$

The first order conditions for c_t^i and $\tilde{X}_t^i$ are

$$c_t^i = \tfrac{1}{\alpha_i + \gamma_i R} \ln \tfrac{\alpha_i}{\beta_i \gamma_i R \delta_i} + \tfrac{\gamma_i R}{\alpha_i + \gamma_i R} \tilde{W}_t^{\,i} + \tfrac{1}{\alpha_i + \gamma_i R} \Phi^i(\tilde{X}_t^i, Z_t), \tag{107}$$

$$\tilde{X}_t^i = -(u^i_{XX})^{-1} u^i_{XZ} Z_t = h^i Z_t. \tag{108}$$

The second order condition is always satisfied for c_t^i, and is satisfied for $\tilde{X}_t^i$ if and only if u^i_{XX} is positive definite (which is always the case, at least in the numerical applications).

For this optimum $\tilde{X}_t^i$ we can calculate

$$\Phi^i(\tilde{X}_t^i, Z_t) = \Phi^i\left(h^i Z_t, Z_t\right) = \tfrac{1}{2} Z_t^\top \left(u^i_{ZZ} - (u^i_{XZ})^\top (u^i_{XX})^{-1} u^i_{XZ}\right) Z_t = \tfrac{1}{2} Z_t^\top g^i Z_t. \tag{109}$$

Substitute (107), (108) and (109) into (106), and identify the coefficients of $\tilde{W}_t^{\,i}$ and Z_t. We

obtain the following equations

$$\gamma_i = \frac{\alpha_i r}{R}, \tag{110}$$

$$v^i = \tfrac{1}{R} g^i - 2 \ln \left(\tfrac{R}{r} \left(\beta_i r \delta_i \right)^{1/R} \right) E^{11}, \tag{111}$$

where E^{11} is the $(n+2) \times (n+2)$ matrix with all entries zero, except for the top left entry, which is one. Now we substitute (110) and (109) into equation (107), and derive the optimal consumption strategy

$$c_t^i = \tfrac{r}{R} \tilde{W}_t{}^i - \tfrac{1}{\alpha_i R} \ln(\beta_i r \delta_i) + \tfrac{1}{\alpha_i R} \tfrac{1}{2} Z_t^\top g^i Z_t. \tag{112}$$

Notice that equations (112) and (108) for $i = N, C$ are exactly the results we stated in Theorems 19 and 20. As we discussed at the beginning of this section, this proves Theorems 18–20, conditional on showing that the equations resulting from verifying our guesses are satisfied by some values of p, q, v^N and v^C. Let us take the guesses in turn.

First, we saw that the equilibrium evolution of Z_t was based on the assumption that the net stock demand of N, x_t, is determined by the formula $x_t = qZ_t$. But x_t is the optimal stock demand $\tilde{X}_t^N$, which has the form $\tilde{X}_t^N = h^N Z_t$. This implies that

$$q = h^N. \tag{113}$$

(Perhaps we should write $h^N(p, q, v^N, v^C)$, to indicate the dependence of h^N on the choice of parameters.)

Second, we assumed that the equilibrium price is $P_t = pZ_t$. Then P_t has to clear the market, which means that the total demand $X_t^N + X_t^C$ has to equal u, the supply vector. We can write $u = f_1 Z_t$ and $X_{t-1} = f_X Z_t$, where $f_u = \begin{bmatrix} u & o & O \end{bmatrix}$ and $f_X = \begin{bmatrix} o & o & I \end{bmatrix}$ are constant $(n+2)$-vectors. So $X_t^N = (h^N + f_X) Z_t$, and from equating supply with demand we get

$$f_u = \left(h^N + f_X \right) + h^C. \tag{114}$$

Finally, we assume that the value functions of the two agents are of the form $J_t^i = -\beta_i^t \exp\left(-\gamma_i \tilde{W}_t{}^i - \tfrac{1}{2} Z_t^\top v^i Z_t \right)$. Then we saw above that applying the Bellman equation (98)

and identifying the coefficients of $\tilde{W}_t^{\,i}$ and Z_t leads to two sets of equations, one for each agent $i = N, C$:

$$v^i = \tfrac{1}{R} g^i - 2 \ln \left(\tfrac{R}{r} \left(\beta_i r \delta_i \right)^{1/R} \right) E^{11}. \tag{115}$$

Putting together identities (113)–(115), we get equations (72), and we are done.

We first prove Proposition 22. In the system of equations (72) we set $\lambda = 0$ and $\sigma_{22} = 0$. We then briefly indicate how to find a solution $\mathbf{x}_0$ of this system, and we show that it is the only solution that leads to a stationary representation for P_t and X_t.

Observe that in (77) the set of equations $\mathbf{F}_{opt^N} = 0$ implies $q = h^N$, which basically expresses q in terms of the other variables: p, v^N and v^C. We could have done this from the beginning and omit $\mathbf{F}_{opt^N} = 0$ from our system, but then the other sets of equations would have looked more complicated. Therefore, we prefer to substitute the components of q in our equations only when needed.

Let us look first at the third equation from the set $\mathbf{F}_{mkt} = 0$ which we denote by $\mathbf{F}_{mkt}(3) = 0$. If we substitute for q_X as discussed above, $\mathbf{F}_{mkt}(3) = 0$ is a second degree equation in p_X, which has the following two real solutions:

$$p_X = 0 \quad \text{and} \quad p_X = \frac{v_{33}^N - \sigma_{11}\gamma_N(\gamma_N + \gamma_C)}{\gamma_N}$$

We first analyze the second solution for p_X, and we show that it leads to a non-stationary representation for P_t and X_t: Notice that the second solution for p_X depends on the variable v_{33}^N, so we are not done yet in finding p_X. But it gives us the idea of looking at the equation $\mathbf{F}_{val^N}(3,3) = 0$. We then substitute q_X and p_X into $\mathbf{F}_{val^N}(3,3) = 0$. We obtain a quadratic equation in v_{33}^N, which leads to two solutions for v_{33}^N. When we substitute them into the expression for p_X, we get the following two solutions for p_X (this time, the expression depends only on the parameters):

$$p_X = -\frac{\sigma_{11}(\gamma_N + \delta)}{2r}, \quad \text{where} \quad \delta = \pm \Big((\gamma_N + 2\gamma_C)^2 - 4R\gamma_C(\gamma_N + \gamma_C) \Big)^{1/2}.$$

Correspondingly, we get two solutions for q_X, corresponding to the two values for δ:

$$q_X = \frac{r(\gamma_N + 2\gamma_C + \delta)}{\gamma_N + 2\gamma_C - 2R\gamma_C + \delta}.$$

Now one can show that either q_X is complex imaginary, or, if it is real, q_X must be positive: for this it is enough to show that when δ is real the denominator $\gamma_N + 2\gamma_C - 2R\gamma_C + \delta$ is positive, or equivalently that $(\gamma_N + 2\gamma_C - 2R\gamma_C)^2 > (\gamma_N + 2\gamma_C)^2 - 4R\gamma_C(\gamma_N + \gamma_C)$. But this is equivalent to $R > 1$, which we know is true.

Now recall that we had the following formula for X_t:

$$X_t = q_1 + q_A A_t + (q_X + 1)X_t.$$

This is the expression of a stationary time series X_t if and only if $q_X + 1 \in (-1, 1)$. But in our case $q_X > 0$, hence $q_X + 1 > 1$, so the resulting expression for X_t is not stationary. We should point out that even though the expression for X_t is non-stationary, the resulting time series X_t might be, and in our case it actually is (recall that $X_t = \omega$, which is a constant time series). However, when we let the parameters λ and σ_{22} be non-zero, the actual time series X_t and P_t become non-stationary, not just their expressions.

That leaves us with the only viable solution corresponding to $p_X = 0$. As we did with the other solution, we look at $\mathbf{F}_{val^N}(3,3) = 0$, and we get $v_{33}^N = 0$. We substitute the values for p_X and v_{33}^N into the formula for q_X and get $q_X = -1$. Since from here on calculations are straightforward, we only briefly indicate how to find the other components of $\mathbf{x}_0$: The equation $\mathbf{F}_{val^C}(3,3) = 0$ implies $v_{33}^C = 0$ (after substituting the values we already found); from $\mathbf{F}_{val^C}(2,3) = 0$ we get $v_{23}^C = 0$; the equation $\mathbf{F}_{val^N}(2,3) = 0$ implies $v_{23}^N = \gamma_C p_A$; using this and $\mathbf{F}_{mkt}(2) = 0$, we obtain p_A, hence also v_{23}^N; next we get q_A, v_{22}^N, v_{22}^C, v_{13}^C; from $\mathbf{F}_{val^N}(1,3) = 0$ we get $v_{13}^N = \gamma_N p_1$; we then obtain p_1 from $\mathbf{F}_{mkt}(1) = 0$ and this determines v_{13}^N; finally, we get v_{12}^N, v_{12}^C, v_{11}^N, v_{11}^C. This completes the proof of Proposition 22.

We now prove Proposition 21. Begin with equations (73) and (74), which describe the equilibrium processes $P_t = p_1 + p_A A_t + p_X X_{t-1}$ and $X_t = q_1 + q_A A_t + q'_X X_{t-1}$, with $q'_X = q_X + 1$. Since we are in the degenerate case when $\sigma_{22} = 0$, the amplitude A_t of the shock is zero. Also, the previous proposition implies that $p_X = 0$ and $q'_X = 0$. That shows that $P_t = p_1 = (D - \gamma_0\sigma_{11})/r$, and $X_t = q_1 = \omega$, so we are done.

The proof of Proposition 23 is quite standard, since it fits into the implicit function theorem framework. In our case, we have a system of equations $\mathbf{F}(\mathbf{x}) = 0$, where we

indicate its dependence on the parameters λ and σ_{22} by writing it as

$$\mathbf{F}(\mathbf{x}, \mathbf{p}) = 0, \quad \text{where} \quad \mathbf{p} = \begin{bmatrix} \lambda & \sigma_{22} \end{bmatrix}^\top.$$

We denote the implicit solution of this system as $\mathbf{x} = \mathbf{x}(\mathbf{p})$. When $\mathbf{p} = 0$, Proposition 22 gives a value for $\mathbf{x}_0 = \mathbf{x}(0)$. We want to find the Taylor expansion of $\mathbf{x}(\mathbf{p})$ around $\mathbf{x}_0$.

The solution of this is standard, and we only indicate how to find the first-order Taylor expansion. We know that $\mathbf{F}(\mathbf{x}(\mathbf{p}), \mathbf{p}) = 0$ for all $\mathbf{p}$. Differentiate this expression with respect to $\mathbf{p}$ to obtain:

$$\mathbf{F}_\mathbf{x}\big(\mathbf{x}(\mathbf{p}), \mathbf{p}\big) \cdot \frac{\partial \mathbf{x}}{\partial \mathbf{p}}(\mathbf{p}) + \mathbf{F}_\mathbf{p}\big(\mathbf{x}(\mathbf{p}), \mathbf{p}\big) = 0 \implies \frac{\partial \mathbf{x}}{\partial \mathbf{p}}(0) = -\big(\mathbf{F}_\mathbf{x}(\mathbf{x}_0, 0)\big)^{-1} \mathbf{F}_\mathbf{p}(\mathbf{x}_0, 0).$$

The $(n+2) \times n$ matrix $\frac{\partial \mathbf{x}}{\partial \mathbf{p}}(0)$ represents the first-order coefficient of the Taylor expansion of $\mathbf{x}$ with respect to $\mathbf{p}$. Now recall that we want to calculate the Taylor expansion not of $\mathbf{x}$, but of $\bar{X}$ and $\bar{P}$. However, $\bar{X} = -q_1/q_X$, and $\bar{P} = p_1 + p_X \bar{X}$, so both $\bar{P}$ and $\bar{X}$ are functions of $\mathbf{x}$. To take their Taylor expansion around $\mathbf{p} = 0$ is therefore standard. The results for some of the coefficients are indicated in the statement of the Proposition.

Of particular interest are coefficients a_{22} and a_{112}, which are proportional to some factors t_{22} and t_{112} that we now define. Let us start with t_{22}. Notice that $\sigma_{13}^2 = \sigma_{11}\sigma_{33}\rho^2$, where $\rho \in [-1, 1]$ is the correlation between $\varepsilon_{D,t}$ and $\varepsilon_{F,t}$. Then we can write

$$t_{22} = \sigma_{11}\sigma_{33}(R - \phi)(R + \phi\omega)(\mu - 1), \tag{116}$$
$$\text{where} \quad \mu = \rho^2(1 - \omega)\left(1 + \frac{\omega\big(\phi(1-\phi) - r(R-\phi)\big)}{(R-\phi)(R+\omega\phi)}\right)$$

Using that $\omega \in (0, 1)$, $r > 0$ and $\phi \in [0, 1)$, we show that $\mu < 1$:

$$\begin{aligned} \mu &\leq (1-\omega)\left(1 + \frac{\omega\phi(1-\phi)}{(1-\phi)(1+\omega\phi)}\right) = (1-\omega)\left(1 + \frac{\omega\phi}{1+\omega\phi}\right) \\ &\leq (1-\omega)\left(1 + \frac{\omega}{1+\omega}\right) \leq (1-\omega)(1+\omega) = 1 - \omega^2 < 1. \end{aligned}$$

But $\mu > 1$ is equivalent to $t_{22} < 0$, which is what we wanted to prove.

We now turn to t_{112}, which is given by the formula:

$$t_{112} = t(R) = aR^2 + bR + c, \quad \text{where} \quad a = -(3\omega + 1), \tag{117}$$
$$b = 6\omega^2 + 4\omega + \omega\phi + \phi - 4\omega^2\phi, \quad c = -\omega(1 + 2\omega - \omega\phi)(1 + \phi + \omega - \omega\phi).$$

One can calculate

$$t(1) = -(1 - \omega)(1 - \phi)(1 + \omega + \phi\omega^2 - 2\omega^2)) < 0,$$
$$t'(1) = \omega^2(6 - 4\phi) - \omega(2 - \phi) - (2 - \phi).$$

Since $t(1) < 0$ and the leading coefficient a is negative, it is clear that $t(R)$ is negative at the endpoints of the interval $[1, \infty)$. The only way $t(R)$ can be ever be positive for some value of R in $[1, \infty)$ is if the quadratic polynomial $t(R)$ has a root in $[1, \infty)$. We therefore have to analyze its discriminant $\Delta = \Delta_t$:

$$\Delta = b^2 - 4ac = u\phi^2 + v\phi + w, \quad \text{where } u = 1 + 2\omega - 3\omega^2 + 4\omega^4$$
$$v = -4\omega(1 + \omega)(3\omega^2 - \omega - 1), \quad w = 4\omega(3\omega^3 + \omega^2 - 2\omega - 1).$$

It is easy to see that $t(R)$ has a root $R \in [1, \infty)$ if and only if $t'(1) > 0$ and $\Delta_t \geq 0$. From the formula for $t'(1)$, one can check that

$$t'(1) > 0 \iff \omega > \omega_0(\phi) = \frac{2 - \phi + \sqrt{(2 - \phi)(26 - 17\phi)}}{12 - 8\phi}. \tag{118}$$

Now we analyze when $\Delta \geq 0$. Since Δ is itself a quadratic polynomial in ϕ, its discriminant equals $w^2 - 4uv = 16\omega(1 + 3\omega)(1 + \omega)(1 - \omega)^2(1 + 4\omega + 2\omega^2) > 0$. That shows that $\Delta(\phi)$ always has real roots. We can also calculate

$$\Delta(1) = (1 - \omega)^2(2\omega + 1)^2 > 0, \qquad \Delta'(1) = 2(1 - \omega)(1 + \omega)(1 + 4\omega + 2\omega^2) > 0,$$
$$\Delta(0) = 4\omega(3\omega^2 + \omega^2 - 2\omega - 1), \qquad \Delta'(0) = -4\omega(3\omega^2 + 2\omega^2 - 2\omega - 1).$$

There are two cases:

1. $\Delta(0) \geq 0$. This is the same as $3\omega^2 + \omega^2 - 2\omega - 1 \geq 0$, which implies $3\omega^2 + 2\omega^2 - 2\omega - 1 >$

0, i.e., $\Delta'(0) < 0$. We know that $\Delta(\phi)$ always has real roots, so in this case both roots lie in $[0,1]$. Denote them by $\phi_0(\omega) \leq \phi_1(\omega)$.

2. $\Delta(0) < 0$. In this case there is a unique root in $[0,1]$, which we denote by $\phi_1(\omega)$.

Now $\Delta(0) \geq 0$ is equivalent to $\omega \geq \omega^*$, where ω^* is the unique real root of the third degree polynomial $3\omega^2 + 2\omega^2 - 2\omega - 1$ (we compute $\omega^* \approx 0.8712$). So we get

$$\begin{aligned} &\text{If } \omega \geq \omega^*, \text{ then} \quad \Delta \geq 0 \iff \phi \in [0, \phi_0(\omega)] \cup [\phi_1(\omega), 1], \\ &\text{If } \omega < \omega^*, \text{ then} \quad \Delta \geq 0 \iff \phi \in [\phi_1(\omega), 1]. \end{aligned} \tag{119}$$

Now we put together the two conditions $t'(1) > 0$ and $\Delta \geq 0$. A little analysis shows that the region where $\omega > \omega_0(\phi)$ in the (ϕ, ω)-space is disjoint from the region where $\phi \in [\phi_1(\omega), 1]$, but includes the region where $\omega \geq \omega^*$ and $\phi \in [0, \phi_0(\omega)]$. This implies that

$$t'(1) > 0 \text{ and } \Delta \geq 0 \iff \omega \geq \omega^* \text{ and } \phi \in [0, \phi_0(\omega)]. \tag{120}$$

That means that when $\omega < \omega^* = 0.8712$, the factor t_{112} is negative, which is what we wanted to prove.

The last result we have to prove is Proposition 25. By looking at the Taylor series for p_1, p_X, q_1, q_X, we observe that $p_1 - (D - \gamma_0\sigma_{11})/r = -\omega p_X$ and $q_1 - \omega = -\omega q'_X$. Denote by $b(\lambda) = p_X$ and $a(\lambda) = q'_X$ (we also have to do a numerical check, since we do not have all the coefficients). Then, since $A_t = 0$, we deduce that $P_t = (D - \gamma_0\sigma_{11})/r - \omega b(\lambda) + b(\lambda)X_{t-1}$, and $X_t = \omega - \omega a(\lambda) + a(\lambda)X_{t-1}$. We get

$$\begin{aligned} P_t - \frac{D - \gamma_0\sigma_{11}}{r} &= b(\lambda)(X_{t-1} - \omega), \\ X_t - \omega &= a(\lambda)(X_{t-1} - \omega). \end{aligned}$$

The second equation gives $X_t - \omega = a(\lambda)^{t+1}(X_{-1} - \omega)$, and the first equation leads to $P_t - (D - \gamma_0\sigma_{11})/r = b(\lambda)a(\lambda)^t(X_{-1} - \omega)$, and we are done.

C Tables for Part III

Dividend: $D_t = D + \varepsilon_{D,t}$. Shock to agent N's wealth at t: $A_{t-1} \cdot \varepsilon_{N,t}$, where A_t is $AR(1)$: $A_t = \alpha_A A_{t-1} + \varepsilon_{A,t}$.

Stock price after trading at t: P_t; price before trading: V_t. Stock holdings of agent N: X_t. Net demand of N: x_t. Price impact coefficient: λ, where $P_t = V_t + \lambda \cdot x_t$.

Equilibrium solution: $P_t = p_1 + p_A A_t + p_X X_{t-1}$; $\quad x_t = q_1 + q_A A_t + q_X X_{t-1}$.

Notations: $\varepsilon_t = \begin{bmatrix} \varepsilon_{D,t} & \varepsilon_{A,t} & \varepsilon_{F,t} \end{bmatrix}^\top$. $\Sigma = \mathbf{E}[\varepsilon_t \varepsilon_t^\top]$. Long term average of X_t: $\bar{X} = -q_1/q_X$. Long term average of P_t: $\bar{P} = p_1 + p_X \bar{X}$.

Value of parameters for numerical results when $n = 1$: a_A (variable—see tables) $r = 0.05$ $\beta_i = 0.8$ $\alpha_i = 2$ $\omega = 0.5$ $D = 0$ $R = 1 + r$ $\gamma_i = \frac{\alpha_i r}{R}$ $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$ $\sigma_{12} = \sigma_{23} = 0$ $\sigma_{13} = 0.5$.

We change notation and define the excess share return by $Q_t^d = P_t + D_t - (1 + r)P_{t-1}$ We also define $Q_t = P_t - (1 + r)P_{t-1}$. Since $\varepsilon_{D,t}$ is i.i.d. and uncorrelated to $\varepsilon_{A,t}$, we deduce that Q_t and Q_t^d have the same covariance structure (although, of course Q_t^d has a bigger variance). If X and Y are two random variables, denote by $\rho(X, Y)$ the correlation coefficient between them, i.e., $\mathrm{corr}(X, Y)$. Denote by $\rho_1(Q)$ the first-order autocorrelation of Q_t, i.e., $\rho(Q_t, Q_{t-1})$. Recall that x_t is the amount net trading at t.

In the case of two risky assets, $n = 2$, indices 1 and 2 correspond to the two risky assets. For example, by $\rho_1(Q)_{ij}$ we denote the (i, j)th element of the first-order autocorrelation matrix of Q_t, i.e., $\rho(Q_{i,t}, Q_{j,t-1})$.

We saw in Section 11 that when $n = 1$, Q_t is $ARMA(2, 2)$:

$$Q_t - (q'_X + a_A)Q_{t-1} + a_A q'_X = \theta_0 \eta_t + \theta_1 \eta_{t-1} + \theta_2 \eta_{t-2},$$

where $\theta_0 = q_A$, $\theta_1 = p_X q_A + p_A q'_X - R q_A$, and $\theta_2 = -R(p_X q_A - p_A q'_X)$.

Table 1: Numeric results for one risky asset, with $\phi = 0.9$ and $\alpha_i = 2.85$, $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$, $\sigma_{13} = 0.5$, $\beta_i = 0.8$, $r = 0.05$, $D = 0$

λ	$\bar{X}$	$\bar{P}$	p_1	p_A	p_X	q_1	q_A	q_X	θ_0	θ_1	θ_2
0	0.3432	-1.7828	-1.7828	-0.3527	-0.0000	0.3432	-0.3706	-1.0000	-0.3706	0.3891	-0.0000
0.01	0.3443	-1.7798	-1.7815	-0.3518	0.0050	0.3327	-0.3580	-0.9664	-0.3580	0.3859	-0.0105
0.02	0.3453	-1.7771	-1.7806	-0.3509	0.0102	0.3226	-0.3459	-0.9342	-0.3459	0.3827	-0.0206
0.1	0.3493	-1.7661	-1.7842	-0.3435	0.0519	0.2555	-0.2696	-0.7313	-0.2696	0.3614	-0.0822
0.2	0.3498	-1.7648	-1.7993	-0.3353	0.0987	0.2039	-0.2139	-0.5829	-0.2139	0.3433	-0.1247
1	0.3344	-1.8066	-1.9118	-0.2994	0.3147	0.0944	-0.1015	-0.2823	-0.1015	0.2895	-0.1921
2	0.3133	-1.8639	-2.0071	-0.2776	0.4571	0.0630	-0.0716	-0.2012	-0.0716	0.2642	-0.1985
3	0.2922	-1.9213	-2.0831	-0.2639	0.5540	0.0482	-0.0585	-0.1650	-0.0585	0.2493	-0.1973
4	0.2702	-1.9809	-2.1507	-0.2540	0.6284	0.0388	-0.0508	-0.1434	-0.0508	0.2390	-0.1950
5	0.2464	-2.0455	-2.2152	-0.2465	0.6890	0.0317	-0.0456	-0.1288	-0.0456	0.2312	-0.1925
6	0.2196	-2.1183	-2.2808	-0.2407	0.7401	0.0259	-0.0419	-0.1180	-0.0419	0.2253	-0.1904
7	0.1875	-2.2055	-2.3525	-0.2364	0.7842	0.0206	-0.0392	-0.1097	-0.0392	0.2209	-0.1887
8	0.1445	-2.3220	-2.4409	-0.2335	0.8228	0.0149	-0.0371	-0.1031	-0.0371	0.2178	-0.1878
9	0.0587	-2.5550	-2.6052	-0.2337	0.8562	0.0057	-0.0359	-0.0979	-0.0359	0.2177	-0.1890
9.1	0.0302	-2.6323	-2.6583	-0.2348	0.8589	0.0029	-0.0360	-0.0976	-0.0360	0.2188	-0.1900

Table 1 (cont'd): Numeric results for one risky asset, with $\phi = 0.9$ and $\alpha_i = 2.85$, $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$, $\sigma_{13} = 0.5$, $\beta_i = 0.8$, $r = 0.05$, $D = 0$

λ	$\bar{X}$	$\bar{P}$	$\sigma(P)$	$\bar{Q}$	$\sigma(Q)$	$\rho_1(Q)$	$\bar{V}$	$\sigma(V)$	$\rho_1(V)$	$\rho(x_t, Q_t)$	q'_X
0	0.3432	-1.7828	0.8091	0.0891	0.3730	-0.0388	0.3034	0.2292	0.0022	0.9944	0
0.01	0.3443	-1.7798	0.8109	0.0890	0.3720	-0.0339	0.2927	0.2211	0.0003	0.9933	0.0336
0.02	0.3453	-1.7771	0.8127	0.0889	0.3709	-0.0292	0.2827	0.2136	0.0001	0.9914	0.0658
0.1	0.3493	-1.7661	0.8263	0.0883	0.3628	-0.0001	0.2250	0.1700	0.0370	0.9608	0.2687
0.2	0.3498	-1.7648	0.8397	0.0882	0.3540	0.0207	0.1861	0.1406	0.1061	0.9166	0.4171
1	0.3344	-1.8066	0.8826	0.0903	0.3160	0.0621	0.1066	0.0805	0.3630	0.7532	0.7177
2	0.3133	-1.8639	0.8985	0.0932	0.2932	0.0736	0.0824	0.0623	0.4623	0.6847	0.7988
3	0.2922	-1.9213	0.9061	0.0961	0.2789	0.0791	0.0709	0.0536	0.5112	0.6489	0.8350
4	0.2702	-1.9809	0.9111	0.0990	0.2686	0.0828	0.0637	0.0481	0.5418	0.6258	0.8566
5	0.2464	-2.0455	0.9155	0.1023	0.2608	0.0855	0.0588	0.0444	0.5633	0.6091	0.8712
6	0.2196	-2.1183	0.9201	0.1059	0.2548	0.0876	0.0551	0.0416	0.5794	0.5963	0.8820
7	0.1875	-2.2055	0.9259	0.1103	0.2503	0.0895	0.0523	0.0395	0.5920	0.5861	0.8903
8	0.1445	-2.3220	0.9342	0.1161	0.2473	0.0911	0.0503	0.0380	0.6022	0.5778	0.8969
9	0.0587	-2.5550	0.9520	0.1277	0.2476	0.0925	0.0491	0.0371	0.6102	0.5712	0.9021
9.1	0.0302	-2.6323	0.9580	0.1316	0.2488	0.0926	0.0493	0.0372	0.6108	0.5707	0.9024

Table 2: Numeric results for two risky assets, with $\phi = 0.9$ and $\alpha_i = 3$,

$\lambda = [\lambda_{11}\ 0; 0\ 0]$, $\sigma_{11} = I_2$, $\sigma_{22} = \sigma_{33} = 1$, $\sigma_{13} = [0.5; 0.5]$, $\beta_i = 0.8$, $r = 0.05$, $D = 0$

λ_{11}	$\bar{X}_1$	$\bar{X}_2$	$\bar{P}_1$	$\bar{P}_2$	$\bar{Q}_1$	$\bar{Q}_2$	$\sigma(Q_1)$	$\rho(Q_1, Q_2)$	$\sigma(Q_2)$	$\rho_1(Q)_{11}$	$\rho_1(Q)_{12}$	$\rho_1(Q)_{21}$	$\rho_1(Q)_{22}$
0	0.3648	0.3648	-1.8148	-1.8148	0.0907	0.0907	0.3330	1.0000	0.3330	-0.0388	-0.0388	-0.0388	-0.0388
0.01	0.3660	0.3653	-1.8113	-1.8135	0.0906	0.0907	0.3323	1.0000	0.3328	-0.0342	-0.0341	-0.0390	-0.0388
0.02	0.3672	0.3657	-1.8081	-1.8124	0.0904	0.0906	0.3316	1.0000	0.3327	-0.0299	-0.0295	-0.0392	-0.0388
0.1	0.3730	0.3680	-1.7914	-1.8056	0.0896	0.0903	0.3250	0.9992	0.3318	-0.0027	-0.0013	-0.0406	-0.0388
0.2	0.3762	0.3697	-1.7822	-1.8009	0.0891	0.0900	0.3170	0.9980	0.3311	0.0170	0.0188	-0.0423	-0.0389
1	0.3789	0.3735	-1.7745	-1.7901	0.0887	0.0895	0.2788	0.9910	0.3295	0.0563	0.0545	-0.0498	-0.0396
2	0.3758	0.3744	-1.7834	-1.7875	0.0892	0.0894	0.2541	0.9861	0.3293	0.0671	0.0609	-0.0544	-0.0399
10	0.3464	0.3707	-1.8675	-1.7979	0.0934	0.0899	0.1830	0.9707	0.3331	0.0861	0.0622	-0.0663	-0.0404
20	0.3141	0.3632	-1.9597	-1.8196	0.0980	0.0910	0.1506	0.9620	0.3386	0.0952	0.0595	-0.0711	-0.0404
50	0.2158	0.3349	-2.2405	-1.9002	0.1120	0.0950	0.1130	0.9480	0.3559	0.1112	0.0547	-0.0769	-0.0404
60	0.1747	0.3224	-2.3581	-1.9361	0.1179	0.0968	0.1074	0.9450	0.3627	0.1148	0.0536	-0.0780	-0.0405
70	0.1224	0.3062	-2.5076	-1.9823	0.1254	0.0991	0.1037	0.9425	0.3708	0.1176	0.0526	-0.0791	-0.0405
80	0.0444	0.2820	-2.7302	-2.0516	0.1365	0.1026	0.1023	0.9406	0.3819	0.1194	0.0516	-0.0802	-0.0407
81	0.0337	0.2786	-2.7610	-2.0611	0.1380	0.1031	0.1024	0.9404	0.3833	0.1194	0.0515	-0.0803	-0.0407
82	0.0220	0.2750	-2.7944	-2.0715	0.1397	0.1036	0.1025	0.9403	0.3849	0.1195	0.0514	-0.0805	-0.0408
85	-0.0217	0.2614	-2.9192	-2.1104	0.1460	0.1055	0.1033	0.9399	0.3904	0.1194	0.0510	-0.0809	-0.0409
86	-0.0409	0.2554	-2.9739	-2.1274	0.1487	0.1064	0.1038	0.9399	0.3927	0.1193	0.0509	-0.0811	-0.0409

Table 2 (cont'd): Numeric results for two risky assets, with $\phi = 0.9$ and $\alpha_i = 3$,

$\lambda = [\lambda_{11}\ 0; 0\ 0]$, $\sigma_{11} = I_2$, $\sigma_{22} = \sigma_{33} = 1$, $\sigma_{13} = [0.5; 0.5]$, $\beta_i = 0.8$, $r = 0.05$, $D = 0$

λ_{11}	$\bar{V}_1$	$\bar{V}_2$	$\sigma(V_1)$	$\rho(V_1, V_2)$	$\sigma(V_2)$	$\rho_1(V)_{11}$	$\rho_1(V)_{12}$	$\rho_1(V)_{21}$	$\rho_1(V)_{22}$	$\rho(x_1, Q_1)$	$\rho(x_1, Q_2)$	$\rho(x_2, Q_1)$	$\rho(x_2, Q_2)$
0	0.2502	0.2502	0.1890	1.0000	0.1890	0.0022	0.0022	0.0022	0.0022	0.9944	0.9944	0.9944	0.9944
0.01	0.2416	0.2509	0.1825	0.9983	0.1895	0.0003	0.0002	0.0026	0.0024	0.9933	0.9931	0.9945	0.9945
0.02	0.2336	0.2515	0.1765	0.9934	0.1900	0.0001	0.0002	0.0030	0.0027	0.9914	0.9907	0.9945	0.9945
0.1	0.1867	0.2556	0.1410	0.9026	0.1931	0.0360	0.0376	0.0057	0.0037	0.9610	0.9509	0.9940	0.9948
0.2	0.1544	0.2583	0.1167	0.7823	0.1952	0.1040	0.0948	0.0079	0.0039	0.9166	0.8916	0.9928	0.9949
1	0.0869	0.2628	0.0657	0.4369	0.1985	0.3612	0.1860	0.0141	0.0031	0.7501	0.6614	0.9875	0.9951
2	0.0659	0.2637	0.0498	0.3258	0.1992	0.4617	0.1761	0.0172	0.0028	0.6795	0.5590	0.9843	0.9951
10	0.0323	0.2674	0.0244	0.1679	0.2021	0.6222	0.1189	0.0252	0.0024	0.5528	0.3662	0.9737	0.9948
20	0.0229	0.2718	0.0173	0.1298	0.2053	0.6653	0.0969	0.0288	0.0023	0.5137	0.3048	0.9672	0.9946
50	0.0145	0.2854	0.0109	0.0965	0.2156	0.7048	0.0751	0.0332	0.0023	0.4736	0.2425	0.9561	0.9945
60	0.0134	0.2908	0.0101	0.0918	0.2197	0.7106	0.0718	0.0340	0.0022	0.4671	0.2326	0.9537	0.9945
70	0.0126	0.2974	0.0095	0.0883	0.2247	0.7147	0.0693	0.0346	0.0022	0.4621	0.2252	0.9517	0.9944
80	0.0122	0.3066	0.0092	0.0859	0.2317	0.7175	0.0675	0.0352	0.0022	0.4585	0.2196	0.9503	0.9944
81	0.0122	0.3079	0.0092	0.0857	0.2326	0.7176	0.0674	0.0353	0.0023	0.4582	0.2191	0.9503	0.9944
82	0.0122	0.3092	0.0092	0.0855	0.2336	0.7178	0.0673	0.0353	0.0023	0.4580	0.2187	0.9502	0.9944
85	0.0123	0.3139	0.0093	0.0851	0.2371	0.7181	0.0670	0.0355	0.0023	0.4574	0.2176	0.9501	0.9944
86	0.0123	0.3159	0.0093	0.0851	0.2386	0.7181	0.0669	0.0355	0.0023	0.4573	0.2174	0.9501	0.9944

References

[1] ADMATI, A. AND P. PFLEIDERER (1988), A theory of intraday patterns: Volume and price variability, *Review of Financial Studies* **1**, 3–40.

[2] AIYAGARI, R. AND M. GERTLER (1991), Asset returns with transaction costs and uninsured individual risk, *Journal of Monetary Economics* **27**, 309–331.

[3] AMIHUD, Y. AND H. MENDELSON (1980), Market-making with inventory, *Journal of Financial Economics* **8**, 31–53.

[4] BASAK, S. (1997), Consumption choice and asset pricing with a non-price-taking agent, *Economic Theory* **10**, 437–462.

[5] BERGIN, J. AND B. MACLEOD (1993), Continuous time repeated games, *International Economic Review* **34**, 21–37.

[6] BIAIS, B., P. HILLION AND C. SPATT (1995), An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* **50**, 1655-1689.

[7] BLOOMFIELD, R., M. O'HARA AND G. SAAR (2003), The "make" or "take" decision in an electronic market: evidence on the evolution of liquidity, preprint.

[8] BOUCHAUD, J-P., M. MEZARD AND M. POTTERS (2002), Statistical properties of the stock order books: empirical results and models, *Mathematics ArXiv*, `http://xxx.lanl.gov`, cond-mat/0203511.

[9] BLACK, F. (1971), Towards a fully automated exchange, first part, *Financial Analysts Journal* **27**, 29–34.

[10] BRENNAN, M. AND A. SUBRAHMANYAM (1996), Market microstructure and asset pricing: on the compensation for illiquidity in stock returns, *Journal of Financial Economics* **41**, 441–464.

[11] CHAKRAVARTY, S. AND C. HOLDEN (1995), An integrated model of market and limit orders, *Journal of Financial Intermediation* **4**, 213–241.

[12] COHEN, K., S. MAIER, R. SCHWARTZ AND D. WHITCOMB (1981), Transaction costs, order placement strategy, and existence of the bid-ask spread, *Journal of Political Economy* **89**, 287–305.

[13] CONSTANTINIDES, G. (1986), Capital market equilibrium with transaction costs, *Journal of Political Economy* **94**, 842–862.

[14] DEMSETZ, H. (1968), The cost of transacting, *Quarterly Journal of Economics* **82**, 33–53.

[15] DOMOWITZ, I. AND A. WANG (1994), Auctions as algorithms, *Journal of Economic Dynamics and Control* **18**, 29–60.

[16] DUFFIE, D., N. GARLEANU AND L. PEDERSEN (2001), Valuation in dynamic bargaining markets, Stanford University preprint, September.

[17] DUMAS, B. AND E. LUCIANO (1991), An exact solution to the portfolio choice problem under transaction costs, *Journal of Finance* **46**, 577–595.

[18] EVANS, L. (1998), *Partial Differential Equations*, Graduate Studies in Mathematics, Vol. 19, American Mathematical Society.

[19] FARMER, D., P. PATELLI AND I. ZOVKO (2003), The predictive power of zero intelligence in financial markets, preprint, September.

[20] FOUCAULT, T. (1999), Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial Markets* **2**, 99–134.

[21] FOUCAULT, T., O. KADAN AND E. KANDEL (2003), Limit order book as a market for liquidity, preprint, October.

[22] FUDENBERG, D. AND J. TIROLE (1991), *Game Theory*, MIT Press.

[23] GABAIX, X., P. GOPIKRISHNAN, V. PLEROU AND E. STANLEY (2003), Theory of large fluctuations in stock market activity, MIT preprint, August.

[24] GARMAN, M. (1976), Market microstructure, *Journal of Financial Economics* **3**, 257–275.

[25] GLOSTEN, L. (1994), Is the electronic open limit order book inevitable?, *Journal of Finance* **49**, 1127–1161.

[26] GLOSTEN, L. AND P. MILGROM (1985), Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* **14**, 71–100.

[27] GOETTLER, R., C. PARLOUR AND U. RAJAN (2003), Equilibrium in a dynamic limit order market, Carnegie Mellon University preprint, May.

[28] GROSSMAN, S. AND G. LAROQUE (1990), Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods, *Econometrica* **58**, 25–51.

[29] GROSSMAN, S. AND M. MILLER (1988), Liquidity and market structure, *Journal of Finance* **43**, 617–633.

[30] HANDA, P. AND R. SCHWARTZ (1996), Limit order trading, *Journal of Finance* **51**, 1835–1861.

[31] HARRIS, L. (1998), Optimal dynamic order submission strategies in some stylized trading problems, *Financial Markets, Institutions and Instruments*, Vol. 7, No. 2.

[32] HARRIS, L. AND J. HASBROUCK (1996), Market versus limit orders: the Superdot evidence on order submission strategy, *Journal of and Quantitative Analysis*, **31**, 213–231.

[33] HASBROUCK, J. AND G. SOFIANOS (1993), The trades of market makers: an empirical analysis of NYSE specialists, *Journal of Finance*, **68**, 1565–1593.

[34] HAUSMAN, J., A. LO, AND C. MCKINLAY (1992), An ordered Probit analysis of transaction stock prices, *Journal of Financial Economics* **31**, 319–330.

[35] HEATON, J. AND D. LUCAS (1996), Evaluating the effects of incomplete markets on risk sharing and asset pricing, *Journal of Political Economy* **104**, 443–487.

[36] HOLLIFIELD, B., R. MILLER, P. SANDAS AND J. SLIVE (2002), Liquidity supply and demand in limit order markets, CEPR Discussion Paper No. 3676, December.

[37] HUANG, M. (2003), Liquidity shocks and equilibrium liquidity premia, *Journal of Economic Theory* **109**, 104–129.

[38] HUANG, R. AND H. STOLL (1997), The components of the bid-ask spread: A general approach, *Review of Financial Studies* **10**, 995–1034.

[39] JAIN, P. (2002), Institutional design and liquidity on stock exchanges around the world, working paper, Indiana University at Bloomington.

[40] KIHLSTROM, R. (2000), Monopoly power in dynamic securities markets, working paper, University of Pennsylvania, Wharton School.

[41] KUMAR, P. AND D. SEPPI (1993), Limit orders and market orders with optimizing traders, Carnegie Mellon preprint.

[42] KYLE, A. (1985), Continuous auctions and insider trading, *Econometrica* **53**, 1315–1335.

[43] LO, A., C. MCKINLAY AND J. ZHANG (2001), Econometric models of limit order executions, *Journal of Financial Economics* **65**, 31–71.

[44] LO, A. AND J. WANG (2001), Trading Volume: Implications of an Intertemporal Capital Asset Pricing Model, forthcoming.

[45] O'HARA, M. (1995), *Market Microstructure Theory*, Blackwell.

[46] PARLOUR, C. (1998), Price dynamics in limit order markets, *Review of Financial Studies* **11**, 789–816.

[47] PASTOR, L. AND R. STAMBAUGH (2003), Liquidity risk and expected stock returns, *Journal of Political Economy* **111**, 642–685.

[48] PRESS, W., S. TEUKOLSKY, W. VETTERLING AND B. FLANNERY (1992), *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press.

[49] ROCK, K. (1996), The specialist's order book and price anomalies, working paper.

[50] SEPPI, D. (1997), Liquidity provision with limit orders and a strategic specialist, *Review of Financial Studies* **10**, 103–150.

[51] SIMON, L. AND M. STINCHCOMBE (1989), Extensive form games in continuous time: pure strategies, *Econometrica* **57**, 1171–1214.

[52] VAYANOS, D. (1998), Transaction Costs and Asset Prices: A Dynamic Equilibrium Model, *Review of Financial Studies* **11**, 1–58.

[53] VAYANOS, D. (1999), Strategic Trading and Welfare in a Dynamic Market, *Review of Economic Studies* **66**, 219–254.

[54] VAYANOS, D. (2001), Strategic Trading in a Dynamic Noisy Market, *Journal of Finance* **56**, 131–171.

[55] VAYANOS, D. AND J-L. VILA (1999), Equilibrium interest rate and liquidity premium with transaction costs, *Economic Theory* **13**, 509–539.

[56] VAYANOS, D. AND T. WANG (2003), Search and endogenous concentration of liquidity in asset markets, MIT preprint.

[57] WILSON, R. (1986), Equilibria of bid-ask markets, in: *Arrow and the Ascent of Economic Theory: Essays in Honor of Kenneth J. Arrow, G. Feiwel (ed.)*, Macmillan Press.

[58] WANG, J. (1994), A model of competitive trading volume, *Journal of Political Economy* **102**, 127–168.