Leaning against the wind

Pierre-Olivier Weill*

January 21, 2006

Abstract

During financial disruptions, marketmakers provide liquidity by absorbing external selling pressure. They buy when the pressure is large, accumulate inventories, and sell when the pressure alleviates. This paper studies optimal dynamic liquidity provision in a theoretical market setting with large and temporary selling pressure, and order-execution delays. I show that competitive marketmakers offer the socially optimal amount of liquidity, provided they have access to sufficient capital. If raising capital is costly, this suggests a policy role for lenient central-bank lending during financial disruptions.

Keywords: marketmaking capital, marketmaker inventory management, financial crisis.

*Department of Finance, New York University Stern School of Business, 44 West Fourth Street, Suite 9-190, New York NY 10012, e-mail: pweill@stern.nyu.edu. First version: July 2003. I am deeply indebted to Darrell Duffie and Tom Sargent, for their supervision, their encouragements, many detailed comments and suggestions. I also thank Narayana Kocherlakota for fruitful discussions and suggestions. I benefited from comments by Manuel Amador, Marco Bassetto, Bruno Biais, Vinicius Carrasco, John Y. Campbell, William Fuchs, Xavier Gabaix, Ed Green, Bob Hall, Ali Hortaçsu, Steve Kohlhagen, Arvind Krishnamurthy, Hanno Lustig, Erzo Luttmer, Eva Nagypal, Lasse Heje Pedersen, Esteban Rossi-Hansberg, Tano Santos, Carmit Segal, Stijn Van Nieuwerburgh, François Velde, Tuomo Vuolteenaho, Ivan Werning, Randy Wright, Mark Wright, Bill Zame, Ruilin Zhou, participants of Tom Sargent’s reading group at the University of Chicago, Stanford University 2003 SITE conference, and of seminar at Stanford University, NYU Economics and Stern, UCLA Anderson, Columbia GSB, Harvard University, University of Pennsylvania, University of Michigan Finance, MIT, the University of Minnesota, the University of Chicago, Northwestern University Economics and Kellogg, the University of Texas at Austin, the Federal Reserve Bank of Chicago, the Federal Reserve Bank of Cleveland, UCLA Economics, the Federal Reserve Bank of Atlanta, and TEMA. The financial support of the Kohlhagen Fellowship Fund at Stanford University is gratefully acknowledged. All errors are mine.
1 Introduction

When disruptions subject financial markets to unusually strong selling pressures, NYSE specialists and NASDAQ marketmakers typically *lean against the wind* by absorbing the market’s selling pressure and creating liquidity: they buy large quantity of assets and build up inventories when selling pressure in the market is large, then dispose of those inventories after that selling pressure has subsided.\(^1\) In this paper, I develop a model of optimal dynamic liquidity provision. To explain how much and when liquidity should be provided, I solve for socially optimal liquidity provision. I argue that some features of the socially optimal allocation would be regarded by a policymaker as symptoms of poor liquidity provision. In fact, these symptoms can be consistent with efficiency. I also show that when they can maintain sufficient capital, competitive marketmakers supply the socially optimal amount of liquidity. If capital-market imperfections prevent marketmakers from raising sufficient capital, this suggests a policy role for lenient central-bank lending during financial disruptions.

The model studies the following scenario. In the beginning at time zero, outside investors receive an aggregate shock which lowers their marginal utility for holding assets relative to cash. This creates a sudden need for cash and induces a large selling pressure. Then, randomly over time, each investor recovers from the shock, implying that the initial selling pressure slowly alleviates. This is how I create a stylized representation of a “flight-to-liquidity” (Longstaff (2004)) or a stock-market crash such as that of October 1987. All trades are intermediated by marketmakers who do not derive any utility for holding assets and who are located in a central marketplace which can be viewed, say, as the floor of the New-York Stock Exchange. I assume the asset market can be illiquid in the sense that traders face order-execution delays. Specifically, investors make contact with marketmakers only after delays that are designed to represent, for example, front-end order capture, clearing, and settlement. While one expects such delays to be short in normal times, the Brady (1988) report suggests that they were unusually long and variable during the crash of October 1987. Similarly, during the crash of October 1997, customers complained about “poor or untimely execution from broker dealers” (SEC Staff Legal Bulletin No. 8 of September 9, 1998). Lastly, McAndrews and Potter (2002) and Fleming and Garbade (2002) document payment and transaction delays, due to disruption of the communication network after the terrorist attacks of September 11, 2001.

\(^1\)This behavior reflects one aspect of the U.S. Securities and Exchange Commission (SEC) Rule 11-b on maintaining fair and orderly markets.
In this economic environment, marketmakers offer buyers and sellers quicker exchange, what Demsetz (1968) called “immediacy”. Marketmakers anticipate that after the selling pressure subsides, they will achieve contact with more buyers than sellers, which will allow them then to transfer assets to buyers in two ways. They can either contact additional sellers, which is time-consuming because of execution delays; or they can sell from their own inventories, which can be done much more quickly. Therefore, by accumulating inventories early, when the selling pressure is large, marketmakers mitigate the adverse impact on investors of execution delays.

The socially optimal asset allocation maximizes the sum of investors’ and marketmakers’ intertemporal utility, subject to the order-execution technology. Because agents have quasi-linear utilities, any other asset allocation could be Pareto improved by reallocating assets and making time-zero consumption transfers. The upper panel of Figure 1 shows the socially optimal time path of marketmakers’ inventory. (The associated parameters and modelling assumptions are described in Section 2.) The graph shows that marketmakers accumulate inventories only temporarily, when the selling pressure is large. Moreover, in this example, it is not socially optimal that marketmakers start accumulating inventories at time zero when the pressure is strongest. This suggests that a regulation forcing marketmakers to promptly act as “buyers of last resort” could in fact result in a welfare loss. For example, if the initial preference shock is sufficiently persistent, marketmakers acting as buyers of last resort will end up holding assets for a very long time, which cannot be efficient given that they are not the final holder of the asset. Lastly, when the economy is close to its steady state (interpreted as a “normal time”) marketmakers should effectively act as “matchmakers” who never hold assets but merely buy and re-sell instantly.

If marketmakers maintain sufficient capital, I show that the socially optimal allocation is implemented in a competitive equilibrium, as follows. Investors can buy and sell assets only when they contact marketmakers. Marketmakers compete for the order flow and can trade among each other at each time. The lower panel of Figure 1 shows the equilibrium price path. It jumps down at time zero, then increases, and eventually reaches its steady-state level. A marketmaker finds it optimal to accumulate inventories only temporarily, when the asset price grows at a sufficiently high rate. This growth rate compensates for the time value of the money spent on inventory accumulation, giving a marketmaker just enough incentive to provide liquidity. A marketmaker thus buys early at a low price and sells later at a high price, but competition implies that the present value of her profit is zero.

Ample anecdotal evidence suggests that marketmakers do not maintain sufficient capital
Figure 1: Features of the Competitive Equilibrium. 

(Brady (1988), Greenwald and Stein (1988), Marès (2001), and Greenberg (2003).) I find that if marketmakers do not maintain sufficient capital, then they are not able to purchase as many assets as prescribed by the socially optimal allocation. If capital-market imperfections prevent marketmakers from raising sufficient capital before the crash, lenient central-bank lending during the crash can improve welfare. Recall that during the crash of October 1987, the Federal Reserve lowered the funds rate while encouraging commercial banks to lend to security dealers (Parry (1997), Wigmore (1998).) 

It is often argued that marketmakers should provide liquidity in order to maintain price continuity and to smooth asset price movements. The present paper steps back from such price-smoothing objective and instead studies liquidity provision in terms of the Pareto criterion. The results indicate that Pareto-optimal liquidity provision is consistent with a discrete price decline at the time of the crash. This suggests that requiring marketmakers to maintain price continuity at the time of the crash might result in a welfare loss.

\footnote{For instance, the glossary of \texttt{www.nyse.com} states that NYSE specialists “use their capital to bridge temporary gaps in supply and demand and help reduce price volatility.” See also the NYSE information memo 97-55.}
Related Literature

Liquidity provision in normal times has been analyzed in traditional inventory-based models of marketmaking. Garman (1976), Amihud and Mendelson (1980), and Mildenstein and Schleef (1983) study pricing and inventory management by risk-neutral monopolistic marketmakers receiving buying and selling orders at random arrival times. Stoll (1978), Ho and Stoll (1981), and O’Hara and Oldfield (1986) study risk-averse monopolistic marketmakers and explain the impact of return and order-flow uncertainty on bid-ask spreads. Ho and Stoll (1983) derive some equilibrium results with competitive marketmakers. Because they study inventory management in normal times, all of the above authors assume that supply and demand curves are time-invariant. In contrast, I study the inventory management of competitive marketmakers under unusual market conditions, when the market is subject to a large and temporary selling pressure. In my model, supply and demand are time-varying. With competitive marketmakers receiving orders at random arrival times, traditional models would feature time variation in the cross-sectional distribution of inventories and as a result would lose much of their tractability. The present model shortcuts this difficulty by assuming that, at each time, marketmakers can trade among each other. Moreover, while traditional models specify exogenous supply and demand curves, I derive them from the solutions of investors’ intertemporal utility maximization problems. This explicit treatment of investors’ preferences facilitates welfare analysis. Lastly, the final difference with this literature is that I address the impact of scarce marketmaking capital on marketmakers’ profit and price dynamics.

Grossman and Miller (1988) and Greenwald and Stein (1991) analyze marketmaking during disruptions from a risk-sharing perspective. In their model, both the sellers and the marketmakers enter a Walrasian market in the first time period and they wait for buyers to enter in the second. With or without marketmakers, assets are allocated to buyers in the second period, implying that marketmakers play no role in facilitating trade between the initial sellers and the later buyers. In the present model, by contrast, the social benefit of liquidity provision is to facilitate trade, in that it speeds up the allocation of assets from the initial sellers to the later buyers. Moreover Grossman and Miller study a two-period model, which means that the timing of liquidity provision is effectively exogenous, in that marketmakers buy in the first period and sell in the second. With its richer intertemporal structure, my model sheds light on the optimal timing of liquidity provision. Bernardo and Welch (2004) propose a two-period model of financial-market run, along the lines of Diamond and Dybvig
Their main objective is to explain the cause of a financial crisis. In the run, their myopic marketmakers end up providing too much liquidity, prior to an uncertain aggregate liquidity shock. The objective of the present model is not to explain the cause of a crisis but rather to develop an intertemporal model of marketmakers optimal liquidity provision, after an aggregate liquidity shock.

The impact of trading delays in security markets is studied by dynamic asset-pricing models with search frictions, such as Duffie, Gârleanu, and Pedersen (2004), Weill (2004), Vayanos and Wang (2004), Spulber (1996) and Hall and Rust (2003). The present model builds specifically on the work of Duffie, Gârleanu, and Pedersen (2005). In their model, marketmakers are matchmakers who, by assumption, cannot hold inventory. By studying investment in marketmaking capacity, they focus on liquidity provision in the long run. By contrast, I study liquidity provision in the short run and view marketmaking capacity as a fixed parameter. In the short run, marketmakers provide liquidity by adjusting their inventory positions.

Another related literature studies the equilibrium and socially optimal entry of middlemen in search-and-matching economies (see, among others, Rubinstein and Wolinsky (1987), Li (1998), Shevchenko (2004), and Masters (2004)). The central objective of these papers is to characterize the size of the middlemen sector in a steady-state where the aggregate amount of middlemen’s inventories remains constant over time. The present paper studies intermediation during a financial crisis, when it is arguably reasonable to take the size of the marketmaking sector as given. In the short run, the marketmaking sector can only gain capacity by increasing its capital and aggregate inventories fluctuate over time.

The remainder of this paper is organized as follows. Section 2 describes the economic environment, Section 3 solves for socially optimal dynamic liquidity provision, Section 4 studies the implementation of this optimum in a competitive equilibrium, and introduces borrowing-constrained marketmakers. Section 4.2 discusses policy implications, and Section 6 concludes. The appendix contains the proofs.

2 The Economic Environment

This section describes the economy and introduces the two main assumptions of this paper. First, there is a large and temporary selling pressure. Second, there are order-execution delays.
2.1 Marketmakers and Investors

Time is treated continuously, and runs forever. A probability space \((\Omega, \mathcal{F}, P)\) is fixed, as well as an information filtration \(\{\mathcal{F}_t, t \geq 0\}\) satisfying the usual conditions (Protter (1990)). The economy is populated by a non-atomic continuum of infinitely lived and risk-neutral agents who discount the future at the constant rate \(r > 0\). An agent enjoys the consumption of a non-storable numéraire good called “cash,” with a marginal utility normalized to 1.\(^3\)

There is one asset in positive supply. An agent holding \(q\) units of the asset receives a stochastic utility flow \(\theta(t)q\) per unit of time. Stochastic variations in the marginal utility \(\theta(t)\) capture a broad range of trading motives such as changes in hedging needs, binding borrowing constraints, changes in beliefs, or risk-management rules such as risk limits. There are two types of agents, marketmakers and investors, with a measure one (without loss of generality) of each. Marketmakers and investors differ in their marginal-utility processes \(\{\theta(t), t \geq 0\}\), as follows. A marketmaker has a constant marginal utility \(\theta(t) = 0\) while an investor’s marginal utility is a two-state Markov chain: the high-marginal-utility state is normalized to \(\theta(t) = 1\), and the low-marginal-utility state is \(\theta(t) = 1 - \delta\), for some \(\delta \in (0, 1)\). Investors transit randomly, and pair-wise independently, from low to high marginal utility with intensity\(^4\) \(\gamma_u\), and from high to low marginal utility with intensity \(\gamma_d\).

These independent variations over time in investors’ marginal utilities create gains from trade. A low-marginal-utility investor is willing to sell his asset to a high-marginal-utility investor in exchange for cash. A marketmaker’s zero marginal utility could capture a large exposure to the risk of the market she intermediates. In addition, it implies that in the equilibrium to be described, a marketmaker will not be the final holder of the asset. In particular, a marketmaker would choose to hold assets only because she expects to make some profit by buying and reselling.\(^5\)

Asset Holdings

The asset has \(s \in (0, 1)\) shares outstanding per investor’s capita. Marketmakers can hold any positive quantity of the asset. The time \(t\) asset inventory \(I(t)\) of a representative marketmaker

---

\(^3\)Equivalently, one could assume that agents can borrow and save cash in some “bank account,” at the interest rate \(\bar{r} = r\). Section 4 adopts this alternative formulation.

\(^4\)For instance, if \(\theta(t) = 1 - \delta\), the time \(\inf\{u \geq 0 : \theta(t + u) \neq \theta(t)\}\) until the next switch is exponentially distributed with parameter \(\gamma_u\). The successive switching times are independent.

\(^5\)The results of this paper hold under the weaker assumption that marketmakers’ marginal utility is \(\theta(t) = 1 - \delta_M\), for some holding cost \(\delta_M > \delta\). Proofs are available from the author upon request.
satisfies the short-selling constraint\(^6\)

\[ I(t) \geq 0. \quad (1) \]

An investor also cannot short-sell and, moreover, he cannot hold more than one unit of the asset. This paper restricts attention to allocations in which an investor holds either zero or one unit of the asset. In equilibrium, because an investor has linear utility, he will find it optimal to hold either the maximum quantity of one or the minimum quantity of zero.

An investor’s type is made up of his marginal utility (high “\(h\),” or low “\(\ell\)”) and his ownership status (owner of one unit, “\(o\),” or non-owner, “\(n\)”). The set of investors’ types is \(\mathcal{T} \equiv \{\ell o, h n, h o, \ell n\}\). In anticipation of their equilibrium behavior, low-marginal-utility owners (\(\ell o\)) are named “sellers,” and high-marginal-utility non-owners (\(h n\)) are “buyers.” For each \(\sigma \in \mathcal{T}\), \(\mu_\sigma(t)\) denotes the fraction of type-\(\sigma\) investors in the total population of investors. These fractions must satisfy two accounting identities. First, of course,

\[ \mu_{\ell o}(t) + \mu_{hn}(t) + \mu_{\ell n}(t) + \mu_{ho}(t) = 1. \quad (2) \]

Second, the assets are held either by investors or marketmakers, so

\[ \mu_{ho}(t) + \mu_{lo}(t) + I(t) = s. \quad (3) \]

### 2.2 Crash and Recovery

I select initial conditions representing the strong selling pressure of a financial disruption. Namely, it is assumed that, at time zero, all investors are in the low-marginal-utility state (see Table 1). Then, as earlier specified, investors transit to the high-marginal-utility state. Under suitable measurability requirements (see Sun (2000), Theorem C), the Law of Large Numbers applies, and the fraction \(\mu_h(t) \equiv \mu_{ha}(t) + \mu_{hn}(t)\) of high-marginal-utility investors solves the ordinary differential equation (ODE)

\[ \dot{\mu}_h(t) = \gamma_u(\mu_{lo}(t) + \mu_{\ell n}(t)) - \gamma_d(\mu_{ho}(t) + \mu_{hn}(t)) \]

\[ = \gamma_u(1 - \mu_h(t)) - \gamma_d\mu_h(t) \]

\[ = \gamma_u - \gamma\mu_h(t), \]

where \(\dot{\mu}_h(t) = d\mu_h(t)/dt\) and \(\gamma \equiv \gamma_u + \gamma_d\). The first term in (4) is the rate of flow of low-marginal-utility investors transiting to the high-marginal-utility state, while the second term

---

\(^6\)One can show that, in the present setup, choosing \(I(t) \geq 0\) is optimal as long as marketmakers incur a flow cost \(c > 1 - \delta\gamma_d/(\rho + \gamma_u + \gamma_d)\) per unit of negative inventory. This cost could capture, for example, the fact that short positions typically involve larger transaction costs and are more risky than long positions.
is the rate of flow of high-marginal-utility investors transiting to the low-marginal-utility state. With the initial condition \( \mu_h(0) = 0 \), the solution of (4) is

\[
\mu_h(t) = y \left(1 - e^{-\gamma t}\right),
\]

(5)

where \( y \equiv \gamma_u/\gamma \) is the steady-state fraction of high-marginal-utility investors. Importantly for the remainder of the paper, it is assumed that

\[
s < y.
\]

(6)

In other words, in steady state, the fraction \( y \) of high-marginal-utility investors exceeds the asset supply \( s \). This will ensure that, asymptotically in equilibrium, the selling pressure has fully alleviated. Figure 2 plots the time dynamic of \( \mu_h(t) \), for some parameter values that satisfy (6). On the Figure, the unit of time is one hour. Years are converted into hours assuming 250 trading days per year, and 10 hours of trading per days. The parameter values used for all of the illustrative computations of this paper, are in Table 2.

Table 1: Initial conditions.

<table>
<thead>
<tr>
<th>( \mu_{lo}(0) )</th>
<th>( \mu_{hn}(0) )</th>
<th>( \mu_{ln}(0) )</th>
<th>( \mu_{ho}(0) )</th>
<th>( I(0) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( s )</td>
<td>0</td>
<td>1 - ( s )</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 2: Dynamic of \( \mu_h(t) \).
Table 2: Parameter Values.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure of Shares</td>
<td>$s$</td>
</tr>
<tr>
<td>Discount Rate</td>
<td>$r$</td>
</tr>
<tr>
<td>Contact Intensity</td>
<td>$\rho$</td>
</tr>
<tr>
<td>Intensity of Switch to High</td>
<td>$\gamma_u$</td>
</tr>
<tr>
<td>Intensity of Switch to Low</td>
<td>$\gamma_d$</td>
</tr>
<tr>
<td>Low marginal utility</td>
<td>$1 - \delta$</td>
</tr>
</tbody>
</table>

Time is measured in years. Assuming that the stock market opens 250 days a year, $\rho = 1000$ means that it takes 2.5 hours to execute an order, on average. The parameter $\gamma = \gamma_u + \gamma_d$ measures the speed of the recovery. Specifically, with $\gamma = 100$, $\mu_h(t)$ reaches half of its steady-state level in about 1.73 days.

2.3 Order-execution delays

This paper departs from the traditional Walrasian model by assuming that the asset market is illiquid, in that there are order-execution delays. Marketmakers intermediate all trades from a central marketplace which can be viewed, say, as the floor of the New York Stock exchange. The asset market is illiquid in the sense that investors cannot contact that marketplace instantly. Instead, an investor establishes contact with marketmakers at Poisson arrival times with intensity $\rho > 0$. Contact times are pairwise independent across investors and independent of marginal utility processes. Therefore, an application of the Law of Large Numbers (under the technical conditions mentioned earlier) implies that contacts between type-$\sigma$ investors and marketmakers occur at a total (almost sure) rate of $\rho \mu_\sigma(t)$. Hence, in a market equilibrium, $\rho \mu_\sigma(t)$ represents the order-flow rate originating from type-$\sigma$ investors.\(^7\)

The random contact times represent a broad range of execution delays, including the time to contact a marketmaker, to negotiate and process an order, to deliver an asset, or to transfer a payment. The parameter $\rho$ is viewed as a measure of marketmaking capacity, encompassing for instance the communication network and the infrastructures needed to execute transactions. One might argue that execution delays are usually quite short and perhaps therefore of little consequence to the quality of an allocation. The Brady (1988)

\(^7\)An alternative specification would let the instantaneous rate of contact with type-$\sigma$ investors be some increasing and strictly concave function of $\mu_\sigma(t)$. This would capture congestions, as the average contact times would be decreasing in $\mu_\sigma(t)$. This alternative non-linear specification is however much less tractable and requires a numerical solution method. Moreover, one might expect that the basic intuitions for the welfare improving role of marketmakers’ liquidity provision would be the same as with the present linear specification.
report shows, however, that during the October 1987 crash, delays were much longer and much more variable than in normal times. In particular, the report documents that many delays were caused by failures of overloaded execution systems, by congestions in the communication network, and by automated protection features. The report suggests that such delays might have amplified liquidity problems in a far-from-negligible manner.\(^8\)

Delays also occurred during the crash of October 1997. The SEC reported that “broker-dealers web servers had reached their maximum capacity to handle simultaneous users” and “telephone lines were overwhelmed with callers who were frustrated by the inability to access information online.” As a result of these capacity problems, customers could not be “routed to their designated market center for execution on a timely basis” and “a number of broker dealers were forced to manually execute some customers orders.”\(^9\) This suggests that technological improvements which followed the 1987 crash did not prevent substantial order-execution delays from arising during the crash of 1997.

### 3 Optimal Dynamic Liquidity Provision

The first objective of this section is to explain the benefit of liquidity provision, addressing how much and when liquidity should be provided. Its second objective is to establish a benchmark against which to judge the market equilibria studied in Sections 4 and 4.2. To these ends, I temporarily abstract from marketmakers’ incentives to provide liquidity and solve for socially optimal allocations, maximizing the sum of investors and marketmakers’ intertemporal utility, subject to order-execution delays. The optimal allocation is found to resemble “leaning against the wind.” Namely, it is socially optimal that a marketmaker accumulates inventories when the selling pressure is strong.

#### 3.1 Asset Allocations

At each time, a representative marketmaker can transfer assets only to her own account or among those of investors who are currently contacting her. For instance, the flow rate \(u_t(t)\) of assets that a marketmaker takes from low-marginal-utility investors is subject to the

---

\(^8\)After describing the selling pressure originating from portfolio insurers, the report notes: “Transaction systems, such as DOT, or market stabilizing mechanisms, such as NYSE specialists, are bound to be crushed by the pressure, however they are designed or capitalized.” See also Wigmore (1998).

order-flow constraint

\[-\rho \mu_{\ell n}(t) \leq u_{\ell}(t) \leq \rho \mu_{\ell o}(t).\] (7)

The upper (lower) bound shown in (7) is the flow of \(\ell o\) (\(\ell n\)) investors who establish contact with marketmakers at time \(t\). Similarly, the flow \(u_h(t)\) of assets that a marketmaker transfers to high-marginal-utility investors is subject to the order-flow constraint

\[-\rho \mu_{ho}(t) \leq u_h(t) \leq \rho \mu_{hn}(t).\] (8)

When the two flows \(u_{\ell}(t)\) and \(u_h(t)\) are equal, a marketmaker is a matchmaker, in the sense that she takes assets from some \(\ell o\) investors (sellers) and transfers them instantly to some \(hn\) investors (buyers). If the two flows are not equal, a marketmaker is not only matching buyers and sellers, but she is also changing her inventory position. For example, if both \(u_{\ell}(t)\) and \(u_h(t)\) are positive, a marketmaker is matching sellers and buyers at the rate \(\min\{u_{\ell}(t), u_h(t)\}\). The net flow \(u_{\ell}(t) - u_h(t)\) represents the rate of change of a marketmaker’s inventory, in that

\[\dot{I}(t) = u_{\ell}(t) - u_h(t).\] (9)

Similarly, the rate of change of the fraction \(\mu_{\ell o}(t)\) of low-marginal-utility owners is

\[\dot{\mu}_{\ell o}(t) = -u_{\ell}(t) - \gamma_u\mu_{\ell o}(t) + \gamma_d\mu_{ho}(t),\] (10)

where the terms \(\gamma_u\mu_{\ell o}(t)\) and \(\gamma_d\mu_{ho}(t)\) reflect transitions of investors from low to high marginal utility, and from high to low marginal utility, respectively. Likewise, the rate of change of the fractions of \(hn\), \(\ell n\), and \(ho\) investors are, respectively,

\[\dot{\mu}_{hn}(t) = -u_h(t) - \gamma_d\mu_{hn}(t) + \gamma_u\mu_{\ell n}(t)\] (11)

\[\dot{\mu}_{\ell n}(t) = u_{\ell}(t) - \gamma_u\mu_{\ell n}(t) + \gamma_d\mu_{hn}(t)\] (12)

\[\dot{\mu}_{ho}(t) = u_h(t) - \gamma_d\mu_{ho}(t) + \gamma_u\mu_{\ell o}(t).\] (13)

**Definition 1 (Feasible Allocation).** A feasible allocation is some distribution \(\mu(t) \equiv (\mu_{\sigma}(t))_{\sigma \in \Sigma}\) of types, some inventory holding \(I(t)\), and some piecewise continuous asset flows \(u(t) \equiv (u_h(t), u_{\ell}(t))\) such that

(i) At each time, the short-selling constraint (1) and the order-flow constraints (7)-(8) are satisfied.

---

12
(ii) The ODEs (9)-(13) hold.

(iii) The initial conditions of Table 1 hold.

Since $u(t)$ is piecewise continuous, $\mu(t)$ and $I(t)$ are piecewise continuously differentiable.

A feasible allocation is said to be constrained Pareto optimal if it cannot be Pareto improved by choosing another feasible allocation and making time-zero cash transfers. As it is standard with quasi-linear preferences, it can be shown that a constrained Pareto optimal allocation must maximize

$$
\int_0^{+\infty} e^{-rt} \left( \mu_{ho}(t) + (1 - \delta)\mu_{lo}(t) \right) dt,
$$

(14)

the equally weighted sum of investors’ intertemporal utilities for holding assets.\(^{10}\) This criterion is deterministic, reflecting pairwise independence of investors’ marginal-utility and contact-time processes. Conversely, an asset allocation maximizing (14) is constrained Pareto optimal. This discussion motivates the following definition of an optimal allocation.

**Definition 2 (Socially Optimal Allocation).** A socially optimal allocation is some feasible allocation maximizing (14).

### 3.2 The Cost and Benefit of Liquidity Provision

This subsection illustrates the social benefits of accumulating inventories. Namely, it considers the no-inventory allocation ($I(t) = 0$, at each time), and shows that it can be improved if marketmakers accumulate a small amount of inventory, when the selling pressure is strong. I start by describing some features of the no-inventory allocation. Substituting $I(t) = 0$ into equation (3) gives

$$
\mu_{lo}(t) = s - \mu_h(t) + \mu_{hn}(t).
$$

(15)

The “crossing time” is the time $t_s$ at which $\mu_h(t_s) = s$. This is, as Figure 2 illustrates, the time at which the fraction $\mu_h(t)$ of high-marginal-utility investors crosses the supply $s$ of assets. Because $\mu_h(t)$ is increasing, equation (15) implies that

$$
\rho \mu_{hn}(t) < \rho \mu_{lo}(t)
$$

(16)

\(^{10}\)Marketmakers intertemporal utility for holding assets is equal to zero and hence does not appear in (14). If a marketmaker marginal utility for holding asset is $(1 - \delta_M) > 0$, then one has to add a term $\int_0^{\infty} e^{-rt}(1 - \delta_M)I(t) dt$ to the above criterion.
if and only if $t < t_s$. Therefore, in the no-inventory allocation, before the crossing time, the selling pressure is “positive,” meaning that marketmakers are in contact with more sellers ($lo$) than buyers ($hn$). After the crossing time, they are in contact with more buyers than sellers.

Intuitively, the no-inventory allocation can be improved as follows. A marketmaker can take an additional asset from a seller before the crossing time, say at $t_1 = t_s - \varepsilon$, and transfer it to some buyer after the crossing time, at $t_2 = t_s + \varepsilon$. Because the transfer occurs around the crossing time, the transfer time $2\varepsilon$ can be made arbitrarily small.

The benefit is that, for a sufficiently small $\varepsilon$, this asset is allocated almost instantly to some high-marginal-utility investor. Without the transfer, by contrast, this asset would continue to be held by a low-marginal-utility investor until either i) the seller transits to a high marginal utility with intensity $\gamma_u$, or ii) the seller establishes another contact with a marketmaker with intensity $\rho$. This means that, without the transfer, this asset would continue to be held by a seller and not by a buyer, with an instantaneous utility cost of $\delta$, incurred for a non-negligible average time of $1/(\gamma_u + \rho)$.

The cost of the transfer is that the asset is temporarily held by a marketmaker and not by a seller, implying an instantaneous utility cost of $1 - \delta$. If $\varepsilon$ is sufficiently small, this cost is incurred for a negligible time and is smaller than the benefit. This intuitive argument can be formalized by studying the following family of feasible allocations.

**Definition 3 (Buffer Allocation).** A buffer allocation is a feasible allocation defined by two times $(t_1, t_2) \in [0, t_s] \times [t_s, +\infty)$, called “breaking times,” such that\(^{11}\)

\[
\begin{align*}
    u_{\ell}(t) &= \rho \mu_{hn}(t) I_{[0, t_1]}(t) + \rho \mu_{lo}(t) I_{[t_1, +\infty)}(t) \\
    u_{h}(t) &= \rho \mu_{hn}(t) I_{[0, t_2]}(t) + \rho \mu_{lo}(t) I_{[t_2, +\infty)}(t) \\
    I(t_2) &= 0.
\end{align*}
\]

The no-inventory allocation is the buffer allocation for which $t_1 = t_2 = t_s$. A buffer allocation has the “bang-bang” property: at each time, either $u_{\ell}(t) = \rho \mu_{lo}(t)$ or $u_{h}(t) = \rho \mu_{hn}(t)$. Because of the linear objective (14), it is natural to guess that a socially optimal allocation will also have this bang-bang property. In the next subsection, Theorem 1 will confirm this conjecture, showing that the socially optimal allocation belongs to the family of buffer allocations.

\(^{11}\)In what follows, $I_A(\cdot)$ denotes the indicator function of some set $A \subseteq \mathbb{R}$.
In a buffer allocation, a marketmaker acts as a “buffer,” in that she accumulates assets when the selling pressure is strong and unwinds these trades when the pressure alleviates. Specifically, as illustrated in Figure 3, a buffer allocation \((t_1, t_2)\) has three phases. In the first phase, when \(t \in [0, t_1]\), a marketmaker does not accumulate inventory \((u_\ell(t) = u_h(t)\) and \(I(t) = 0\)). In the second phase, when \(t \in (t_1, t_2)\), a marketmaker first builds up \((u_\ell(t) > u_h(t)\) and \(I(t) > 0\)) and then unwinds \((u_\ell(t) < u_h(t)\) and \(I(t) > 0\)) her inventory position. At time \(t_2\), her inventory position reaches zero. In the third phase \(t \in [t_2, +\infty)\), a marketmaker does not accumulate inventory \((u_\ell(t) = u_h(t)\) and \(I(t) = 0\)). The following proposition characterizes buffer allocations by the maximum inventory position held by marketmakers.

**Proposition 1.** There exist some \(\bar{m} \in \mathbb{R}_+\), some strictly decreasing continuous function \(\psi : [0, \bar{m}] \to \mathbb{R}_+\), and some strictly increasing continuous functions \(\phi_i : [0, \bar{m}] \to \mathbb{R}_+, i \in \{1, 2\}\), such that, for all \(m \in [0, \bar{m}]\) and all buffer allocations \((t_1, t_2)\),

\[
    m = \max_{t \in \mathbb{R}_+} I(t) 
\]

\[
    \psi(m) = \arg \max_{t \in \mathbb{R}_+} I(t) 
\]

\[
    t_1 = \psi(m) - \phi_1(m) 
\]

\[
    t_2 = \psi(m) + \phi_2(m), 
\]

where \(\bar{m}\) is the unique solution of \(\psi(z) - \phi_1(z) = 0\). Furthermore, \(\psi(0) = t_s\) and \(\phi_1(0) = \phi_2(0) = 0\).

In words, the breaking times \((t_1, t_2)\) of a buffer allocation can be written as functions of the maximum inventory position \(m\). The maximum inventory position is achieved at time \(\psi(m)\).
In addition, the larger is a marketmaker’s maximum inventory position, the earlier she starts to accumulate and the longer she accumulates. Lastly, if she starts to accumulate at time zero, then her maximum inventory position is $\bar{m}$.

The social welfare (14) associated with a buffer allocation can be written as $W(m)$, for some function $W(\cdot)$ of the maximum inventory position $m$. As anticipated by the intuitive argument, one can prove the following result.

**Proposition 2.**

$$
\lim_{m \to 0^+} \frac{W(m) - W(0)}{m} > 0.
$$

(24)

This demonstrates that the no-inventory allocation ($m = 0$) is improved by accumulating a small amount of inventory near the crossing time $t_s$.

Having shown that accumulating some inventory improves welfare, one would like to explain how much inventory marketmakers should accumulate. Some intuition on this issue can be gained with the following numerical computations. (Theorem 1 will provide the exact answer). For a given buffer allocation ($\mu^m(t), I^m(t), u^m(t)$) with maximum inventory position $m$, I define the cost of holding inventory as

$$
C(m) \equiv \int_0^{+\infty} e^{-rt}(1 - \delta)I^m(t) \, dt,
$$

(25)

the intertemporal utility which is lost because some assets are temporarily held by marketmakers rather than by sellers. Figure 4 shows a numerical computation of $C(m)$. The convexity suggests that the marginal cost of holding inventory is increasing in the maximum inventory position. Intuitively, an additional unit of inventory is transferred later in time,
implying that the holding cost $1 - \delta$ is incurred for a longer time period. Similarly, the benefit of holding inventory is implicitly defined by

$$W(m) \equiv B(m) - C(m).$$

(26)

That is, $B(m) = W(m) + C(m)$ is a measure of social welfare which is compensated for the holding cost $1 - \delta$ of a marketmaker. Figure 5 shows a numerical computation of $B(m)$. The concavity suggests that the marginal benefit of accumulating inventory is decreasing in the maximum inventory position: an additional unit of inventory is transferred to a buyer later in time, which represents a smaller benefit because agents are impatient. Interestingly, $B(\cdot)$ decreases above some inventory level. In this decreasing branch, marketmakers take too long to transfer a marginal unit. It would be faster, on average, to simply wait for the low investors to transit to the high-marginal-utility state.

Overall, these computations suggest that providing liquidity is cheap and valuable near the crossing time ($m$ close to zero). By contrast, providing liquidity near time zero, when the selling pressure is strongest ($m$ close to $\bar{m}$), is both more expensive and less valuable. The marginal social value of providing liquidity near time zero can even be negative, as illustrated by Figures 4 and 5.

3.3 The Socially Optimal Allocation

This subsection provides first-order sufficient conditions for, and solves for, a socially optimal allocation. The reader may wish to skip the following paragraph on first-order conditions, and go directly to Theorem 1, which describes the socially optimal allocation.
First-Order Sufficient Conditions

The first-order sufficient conditions are based on Seierstad and Sydsæter (1977). The accounting identities \( \mu_{\ell o}(t) = \mu_h(t) - \mu_{hn}(t) \) and \( \mu_{\ell n}(t) = 1 - \mu_h(t) - \mu_{\ell n}(t) \) are substituted into the objective and the constraints, reducing the state variables to \( (\mu_{\ell o}(t), \mu_{hn}(t), I(t)) \). The “current-value” Lagrangian (see Kamien and Schwartz (1991), Part II, Section 8) is

\[
\mathcal{L}(t) = \mu_h(t) - \mu_{hn}(t) + (1 - \delta)\mu_{\ell o}(t) + \lambda_\ell(t) \left( -u_\ell(t) - \gamma_a\mu_{\ell o}(t) - \gamma_d\mu_{hn}(t) + \gamma_d\mu_h(t) \right) - \lambda_h(t) \left( -u_h(t) - \gamma_a\mu_{\ell o}(t) - \gamma_d\mu_{hn}(t) + \gamma_a(1 - \mu_h(t)) \right) + \lambda_I(t) \left( u_\ell(t) - u_h(t) \right) + w_\ell(t) \left( \rho\mu_{\ell o}(t) - u_\ell(t) \right) + w_h(t) \left( \rho\mu_{hn}(t) - u_h(t) \right) + \eta_I(t) I(t). \tag{27}
\]

The multiplier \( \lambda_\ell(t) \) of the ODE (10) represents the social value of increasing the flow of investors from the \( \ell n \) type to the \( \ell o \) type or, equivalently, the value of transferring an asset to an \( \ell n \) investor. One gives a similar interpretation to the multipliers \( \lambda_h(t) \) and \( \lambda_I(t) \) of the ODEs (11) and (9), respectively.\footnote{In equation (27), the minus sign in front of \( \lambda_h(t) \) is contrary to conventional notations but turns out to simplify the exposition.} The multipliers \( w_\ell(t) \) and \( w_h(t) \) of the flow constraints (7) and (8) represent the social value of increasing the rate of contact with \( \ell o \) and \( hn \) investors, respectively.\footnote{It is anticipated that the left-hand constraints in (7) and (8) never bind. In other words, a marketmaker never transfers asset from a high-marginal-utility to a low-marginal-utility investor.} The multiplier on the short-selling constraint (1) is \( \eta_I(t) \). The first-order condition with respect to the controls \( u_\ell(t) \) and \( u_h(t) \) are

\[
w_\ell(t) = \lambda_I(t) - \lambda_\ell(t) \tag{28}
\]

\[
w_h(t) = \lambda_h(t) - \lambda_I(t), \tag{29}
\]

respectively. For instance, (28) decomposes \( w_\ell(t) \) into the opportunity cost \(-\lambda_\ell(t)\) of taking assets from \( \ell o \) investors, and the benefit \( \lambda_I(t) \) of increasing a marketmaker’s inventory. The positivity and complementary-slackness conditions for \( w_\ell(t) \) and \( w_h(t) \), respectively, are

\[
w_\ell(t) \geq 0 \quad \text{and} \quad w_\ell(t) \left( \rho\mu_{\ell o}(t) - u_\ell(t) \right) = 0, \tag{30}
\]

\[
w_h(t) \geq 0 \quad \text{and} \quad w_h(t) \left( \rho\mu_{hn}(t) - u_h(t) \right) = 0. \tag{31}
\]

The multipliers \( w_\ell(t) \) and \( w_h(t) \) are non-negative because a marketmaker can ignore additional contacts. The complementary-slackness condition (30) means that, when the marginal
value \(w(t)\) of additional contact is strictly positive, a marketmaker should take the assets of all \(\ell\) investors with whom she is currently in contact. One also has the positivity and complementary-slackness conditions

\[
\eta_I(t) \geq 0 \quad \text{and} \quad \eta_I(t) I(t) = 0. \tag{32}
\]

The ODE for the the multipliers \(\lambda_\ell(t), \lambda_h(t), \text{ and } \lambda_I(t)\) are

\[
\begin{align*}
    r\lambda_\ell(t) &= 1 - \delta + \gamma_u(\lambda_h(t) - \lambda_\ell(t)) + \rho w_\ell(t) + \dot{\lambda}_\ell(t) \\
    r\lambda_h(t) &= 1 + \gamma_d(\lambda_\ell(t) - \lambda_h(t)) - \rho w_h(t) + \dot{\lambda}_h(t) \\
    r\lambda_I(t) &= \eta_I(t) + \dot{\lambda}_I(t),
\end{align*}
\tag{33-35}
\]

respectively. For instance, (33) decomposes the flow value \(r\lambda_\ell(t)\) of transferring an asset to a low-marginal-utility investor. The first term, \(1 - \delta\), is the flow marginal utility of a low-marginal-utility investor holding one unit of the asset. The second term, \(\gamma_u(\lambda_h(t) - \lambda_\ell(t))\), is the expected rate of net utility associated with a transition to high marginal utility. That is, with intensity \(\gamma_u\), \(\lambda_\ell(t)\) becomes the value \(\lambda_h(t)\) of transferring an asset to a high-marginal-utility investor. The third term, \(\rho w_\ell(t)\), is the expected rate of net utility of a contact between an \(\ell\) investor and a marketmaker. The multipliers \((\lambda_\ell(t), \lambda_h(t), \lambda_I(t))\) must satisfy the following additional restrictions. First, they must satisfy the transversality conditions

\[
\lim_{t \to +\infty} \lambda_j(t)e^{-rt} = 0,
\tag{36}
\]

for \(j \in \{\ell, h, I\}\). Second, the multipliers \(\lambda_h(t)\) and \(\lambda_\ell(t)\) are continuous. Because the control variable \(u(t)\) does not appear in the short-selling constraint \(I(t) \geq 0\), however, the multiplier \(\lambda_I(t)\) might jump, with the restriction that

\[
\lambda_I(t+) - \lambda_I(t-) \leq 0 \quad \text{if } I(t) = 0. \tag{37}
\]

In other words, the multiplier \(\lambda_I(t)\) can jump down, but only when the short-selling constraint is binding. Intuitively, if \(\lambda_I(t)\) were to jump up at \(t\), a marketmaker could accumulate additional inventory shortly before \(t\), say a quantity \(\varepsilon\), improving the objective by

\[
\varepsilon(\lambda_I(t+) - \lambda_I(t-))e^{-rt}. \tag{15}
\]

\[14\text{This condition allows to complete the standard optimality verification argument when the time horizon is infinite.}
\[15\text{Because the ODEs for the state variables are linear, the first-order sufficient conditions impose no sign restriction on the multipliers } \lambda_h(t), \lambda_\ell(t), \text{ and } \lambda_I(t)\text{ (see Section 3 in Part II of } \text{Kamien and Schwartz (1991)}.\]
Appendix B guesses and verifies that the (essentially unique) socially optimal allocation is a buffer allocation. Namely, for a given buffer allocation, one constructs multipliers solving the first-order conditions (28) through (37). The restriction \( w(t_1) = 0 \) is used to find the breaking-times \( t_1 \) and \( t_2 \).

**Theorem 1 (Socially optimal Allocation).** There exists a socially optimal allocation \((\mu^*(t), I^*(t), u^*(t), t \geq 0)\). This allocation is a buffer allocation with breaking times \((t^*_1, t^*_2)\) determined by

\[
e^{-\gamma t^*_1} = \left(1 - \frac{s}{y}\right) \frac{1 - e^{-\rho \Delta^*}}{\rho} \frac{\rho - \gamma}{e^{-\gamma \Delta^*} - e^{-\rho \Delta^*}},
\]
\[
t^*_2 = t^*_1 + \Delta^*,
\]
\[
\Delta^* = \min \left\{ \frac{1}{r + \rho} \log \left(1 + \frac{\delta (r + \rho)}{\gamma_u + (1 - \delta)(r + \rho + \gamma_d)} \right), \bar{\Delta} \right\},
\]

where \( \bar{\Delta} \equiv \phi_1(\bar{m}) + \phi_2(\bar{m}) \) and, if \( \gamma = \rho \), one lets \( (e^{-\gamma x} - e^{-\rho x})/(\rho - \gamma) \equiv x \) for all \( x \in \mathbb{R} \).

If \( \Delta^* = \bar{\Delta} \), the first breaking time is \( t^*_1 = 0 \), meaning that a marketmaker starts accumulating inventory at the time of the “crash.”

The socially optimal allocation has three main features. First, it is optimal that a marketmaker provides some liquidity: From time \( t^*_1 \) to time \( t^*_2 \), she builds up and unwinds a positive inventory position. Second, it is not necessarily optimal that a marketmaker provides liquidity at time zero, when the selling pressure is strongest. This suggests that, although a marketmaker should provide liquidity, she should not act as a “buyer of last resort.” Third, when the economy is close to its steady state, interpreted as a normal time, a marketmaker should act as a mere matchmaker, meaning that she should buy and sell instantly. Thus, the socially optimal allocation draws a sharp distinction between socially optimal marketmaking in a normal time of low selling pressure, versus a bad time of strong selling pressure.

**Proposition 3 (Uniqueness).** If \((\mu(t), I(t), u(t))\) is a socially optimal allocation, then \( \mu(t) = \mu^*(t) \) and \( I(t) = I^*(t) \), for all \( t \in \mathbb{R}_+ \).

### 3.4 Comparative Statics

A natural measure of the amount of liquidity provided by marketmakers is the length \( \Delta^* \) of the inventory-accumulation period. Another measure would be the maximum inventory
position $m^*$, which is strictly monotonic with $\Delta^*$, provided that $\rho$ and $\gamma$ are held constant.\footnote{Specifically, $m^* = (\phi_1 + \phi_2)^{-1}(\Delta^*)$, where $\phi_1(m)$ and $\phi_2(m)$ are increasing in $m$. These two functions, however, implicitly depend on $\rho$ and $\gamma$.} With Theorem 1, $\Delta^*$ can be written as $F(\rho, r, \delta, \gamma_u, \gamma_d)$, for some continuous function $F(\cdot)$. The following proposition provides some natural comparative statics.

**Proposition 4 (A Comparative Static).** Let $x = (\rho, r, \delta, \gamma_u, \gamma_d)$ be a vector of exogenous parameters. If $F(x) < \bar{\Delta}$, then $F(\cdot)$ is differentiable at $x$, with partial derivatives

$$\frac{\partial F}{\partial \rho} < 0, \quad \frac{\partial F}{\partial r} < 0, \quad \frac{\partial F}{\partial \delta} > 0, \quad \frac{\partial F}{\partial \gamma_d} < 0, \quad \text{and} \quad \frac{\partial F}{\partial \gamma_u} < 0. \quad (40)$$

If, on the other hand, $F(x) = \bar{\Delta}$, then marketmakers provide the maximum amount of liquidity, in that they start accumulating inventory at time zero, when the selling pressure is strongest. In that case, locally, $F(\cdot)$ does not depend on $(r, \delta)$. Proposition 4 shows that the inventory-accumulation period is longer when an investor’s holding cost $\delta$ is larger. It is shorter whenever the order-execution technology is faster, the agents are more impatient, or investors’ switching intensities $\gamma_d$ and $\gamma_u$ are larger. This last comparative static reflects the fact that larger $\gamma_d$ and $\gamma_u$ reduce the net utility of transferring the asset. Namely, with a larger $\gamma_d$, an investor keeps a high marginal utility for a shorter time, on average. As a result, the net utility of transferring the asset to an $hn$ investor is smaller. With a larger $\gamma_u$, an $lo$ investor transits faster to a high marginal utility. This increases the value of leaving the asset to this $lo$ investor, and waiting for him to transit to the $ho$ type. Hence, this decreases the net utility of transferring the asset.

**Walrasian Limit**

This paragraph studies the socially optimal allocation as $\rho$ goes to infinity, interpreted as the Walrasian limit with no execution delay. This comparative static exercise illustrates the crucial role of order-execution delays in making it optimal for a marketmaker to provide some amount of liquidity to investors. Specifically, it is shown that, in the limit $\rho \to +\infty$, a marketmaker should not provide any liquidity.

Theorem 1 implies that the length $\Delta^*$ of the inventory-accumulation period goes to zero. This does not immediately imply that the maximum inventory position, $m^* = (\phi_1 + \phi_2)^{-1}(\Delta^*)$, goes to zero. Although marketmakers accumulate inventories during increasingly
small time periods as $\rho \to +\infty$, they also accumulate inventories increasingly quickly. The following proposition settles this issue.

**Proposition 5 (Walrasian Limit).** Given some $(r, \delta, \gamma_u, \gamma_d)$, as $\hat{\rho} \to +\infty$,

\[
F(\hat{\rho}, r, \delta, \gamma_u, \gamma_d) \to 0
\]

\[
(\phi_1 + \phi_2)^{-1} \circ F(\hat{\rho}, r, \delta, \gamma_u, \gamma_d) \to 0.
\]

As the average execution delay $1/\rho$ approaches zero, an asset can be transferred almost instantly to some high-marginal-utility investor. Then, the inventory holding cost is large relative to the benefit of providing liquidity. In such circumstances, a marketmaker should hold a smaller quantity of the asset, for a shorter time.

### 4 Market Equilibrium

This section studies marketmakers’ incentives to provide liquidity. I show that the socially optimal allocation can be implemented in a competitive equilibrium as long as they have access to sufficient capital.

#### 4.1 Competitive Marketmakers

This subsection describes a competitive market structure that implements the socially optimal allocation.

It is assumed that a marketmaker has access to some bank account earning the constant interest rate $\bar{r} = r$. At each time $t$, she buys a flow $u_{\ell}(t) \in \mathbb{R}_+$ of assets, sells a flow $u_h(t) \in \mathbb{R}_+$, and consumes cash at the positive rate $c(t) \in \mathbb{R}_+$. She takes as given the asset price path $\{p(t), t \geq 0\}$. Hence, her bank account position $a(t)$ and her inventory position $I(t)$ evolve according to

\[
\dot{a}(t) = ra(t) + p(t)(u_h(t) - u_{\ell}(t)) - c(t)
\]

\[
\dot{I}(t) = u_{\ell}(t) - u_h(t).
\]

In addition, she faces the borrowing and short-selling constraints

\[
a(t) \geq 0
\]

\[
I(t) \geq 0.
\]
Lastly, at time zero, a marketmaker holds no inventory \((I(0) = 0)\) and maintains a strictly positive amount of capital \(a(0)\). This subsection restricts attention to some large \(a(0)\), in the sense that the borrowing constraint \((45)\) does not bind in equilibrium. (This statement is made precise by Theorem 2 and Proposition 6.) The marketmaker’s objective is to maximize the present value
\[
\int_0^{+\infty} e^{-rt} c(t) \, dt \tag{47}
\]
of her consumption stream with respect to \(\{a(t), I(t), u_\ell(t), u_h(t), c(t), t \geq 0\}\), subject to the constraints \((43)-(46)\), and the constraint that \(u_\ell(t)\) and \(u_h(t)\) are piecewise continuous.\(^{17}\)

Let’s turn to the investor’s problem. An investor establishes contact with some marketmaker at Poisson arrival times with intensity \(\rho > 0\). Conditional on establishing contact at time \(t\), he can buy or sell the asset at price \(p(t)\). I solve the investor’s problem using a “guess and verify” method. Specifically, I guess that, in equilibrium, an \(\ell o\) (\(hn\)) investor always finds it weakly optimal to sell (buy). If an \(\ell o\) (\(hn\)) investor is indifferent between selling and not selling, he might choose not to sell (buy). Lastly, I guess that investors of types \(\ell n\) and \(ho\) never trade. The time-\(t\) continuation utility of an investor of type \(\sigma \in \mathcal{I}\) who follows this policy is denoted \(V_\sigma(t)\). Hence, a seller’s reservation value is \(\Delta V_\ell(t) \equiv V_{\ell o}(t) - V_{\ell n}(t)\), and a buyer’s reservation value is \(\Delta V_h(t) \equiv V_{ho}(t) - V_{hn}(t)\). Appendix C.1 provides ODEs for these continuation utilities and reservation values, as well as a precise definition of a competitive equilibrium. The main result of this subsection states that the optimal allocation can be implemented in some equilibrium:

**Theorem 2 (Implementation).** There exists some \(\overline{a}_0^* \in \mathbb{R}_+\) such that, for all \(a(0) \geq \overline{a}_0^*\), there exists a competitive equilibrium whose allocation is the optimal allocation.

The proof identifies the price and the reservation values with the Lagrange multipliers of the socially optimal allocation (see Table 3). For instance, the asset price \(p(t)\) is equal to the multiplier \(\lambda_\ell(t)\) for the ODE \(\dot{I}(t) = u_\ell(t) - u_h(t)\), interpreted as the social value of increasing the inventory position of a marketmaker. Also, Table 3 and the complementary slackness condition \((30)\) imply that, before the first breaking time \(t^*_1\), \(w_\ell(t) = p(t) - \Delta V_\ell(t) = 0\). In other words, because the socially optimal allocation prescribes that marketmakers do not accommodate the selling pressure, the social value \(w_\ell(t)\) of increasing the rate of contact with

\(^{17}\)The above formulation implies that, in equilibrium, if the borrowing constraint \((45)\) never binds, a marketmaker’s intertemporal utility is equal to \(a(0)\), meaning that equilibrium net profit must be equal to zero.
### Table 3: Identifying Prices with Multipliers.

<table>
<thead>
<tr>
<th>Equilibrium Objects</th>
<th>Multipliers</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(t)$</td>
<td>$\lambda_I(t)$</td>
<td>$\dot{I}(t) = u_\ell(t) - u_h(t)$</td>
</tr>
<tr>
<td>$\Delta V_\ell(t)$</td>
<td>$\lambda_\ell(t)$</td>
<td>$\dot{\mu}<em>{\ell\ell}(t) = -u</em>\ell(t)\ldots$</td>
</tr>
<tr>
<td>$p(t) - \Delta V_\ell(t)$</td>
<td>$w_\ell(t)$</td>
<td>$u_\ell(t) \leq \rho \mu_{\ell\ell}(t)$</td>
</tr>
<tr>
<td>$\Delta V_h(t)$</td>
<td>$\lambda_h(t)$</td>
<td>$\dot{\mu}_{hn}(t) = -u_h(t)\ldots$</td>
</tr>
<tr>
<td>$\Delta V_h(t) - p(t)$</td>
<td>$w_h(t)$</td>
<td>$u_h(t) \leq \rho \mu_{hn}(t)$</td>
</tr>
</tbody>
</table>

In a given row, the equilibrium object in the first column is equal to the Lagrange multiplier in the second column. The third column describes the constraints associated with these multipliers. For instance, in the first row, the price $p(t)$ (first column) is equal to the multiplier $\lambda_I(t)$ (second column) of the ODE $\dot{I}(t) = u_\ell(t) - u_h(t)$ (third column).

sellers is zero. In equilibrium, this means that the price adjusts so that sellers are indifferent between selling and not selling.

**Equilibrium Price Path and Marketmakers’ Incentive**

Appendix B.2 derives closed-form solutions for the equilibrium price path $p(t)$ and the reservation values $\Delta V_\ell(t)$ and $\Delta V_h(t)$. The price path, shown in the lower panel of Figure 6, jumps down at time zero, then increases, and eventually stabilizes at its steady-state level.\(^\text{18}\) The price path reflects the three phases of the socially optimal allocation: before the first breaking time $t^*_1$, sellers are indifferent between selling or not selling, meaning that $p(t) = \Delta V_\ell(t) = \lambda_\ell(t)$. Moreover, equation (33) shows that the growth rate $\dot{p}(t)/p(t)$ of the price is strictly less than $r - (1 - \delta)/p(t)$. This is because a seller with marginal utility $1 - \delta$ does not need a large capital gain to be willing to hold the asset. By contrast, in between the two breaking times $t^*_1$ and $t^*_2$, marketmakers accommodate all of the selling pressure. As a result, the “marginal investor” is a marketmaker and $p(t) > \Delta V_\ell(t)$, meaning that the liquidity provision of marketmakers raises the asset price above a seller’s reservation value. Moreover, equation (35) shows that the price grows at the higher rate $\dot{p}(t)/p(t) = r$, implying

\(^{18}\)A simple way to construct the initial price jump is to start the economy in steady state at $t = 0$ and assume that agents anticipate a crash at some Poisson arrival time with intensity $\kappa$. One can show that the results of this paper would apply, provided that either $\kappa$ is small enough or $t^*_1 > 0$. For Figure 6, it is assumed that $\kappa = 0.$
that the price recovers more quickly when marketmakers provide liquidity.

The capital gain during $[t^*_1, t^*_2]$ exactly compensates a marketmaker for the time value of cash spent on liquidity provision. In other words, a marketmaker is indifferent between i) investing cash in her bank account, and ii) buying assets after $t^*_1$ and selling them before $t^*_2$. Before $t^*_1$ and after $t^*_2$, however, the capital gain is strictly smaller than $r$, making it unprofitable for the marketmaker to buy the asset on her own account. Therefore, in equilibrium, a marketmaker’s intertemporal utility is equal to $a(0)$, the value of her time-zero capital. In other words, although a marketmaker buys low and sells high, competition drives the present value of her profit to zero.

Limit Orders versus Marketmakers

In all of the above, it is assumed that investors cannot trade with limit orders. This might be viewed as a strong assumption, since limit orders would allow sellers to “queue” their assets in the limit order book. In other words, even if sellers are not continuously in contact with the market, their limit orders would be continuously available for trading. Hence, a limit order book might be a natural substitute for marketmakers’ inventory accumulation.

In practice, however, a seller might find limit orders unattractive because of the risk of being “picked off” (see, among others, Copeland and Galai (1983)). Namely, a limit order to sell may be executed when good news causes both the seller’s reservation value and the market price to jump above the limit price. Because, by definition, the order is executed at

---

$^{19}$In that sense, marketmakers’ optimality conditions are similar to those of speculators in commodity markets (see, for example, Flood and Garber (1984)).
the limit price, the seller makes a net loss.\(^{20}\) One might conjecture that, during a market disruption with frequent price jumps, a seller would use limit orders cautiously and would sometimes prefer to trade directly with marketmakers. (Schwert (1998) documents large price increases and declines during financial disruptions.) The evidence of Goldstein and Kavajecz (2004) is consistent with this conjecture. They show that, during the market break of October 1997, there was a dramatic drain of liquidity in the limit order book. Namely, for the Dow Jones Industrial Average stocks of their sample, the limit order book spread could be as high as 3 to 4 dollars. Meanwhile, the quoted spread for the same stocks was about 20 cents. This suggests that sellers (buyers) were mitigating the risk of being picked off by posting very high (low) limit orders.

In Addendum I, I formalizes this discussion by extending the present model in two ways. First, aggregate uncertainty creates stochastic price jumps. Second, an investor who contacts the market has the choice to either trade instantly at the current market price, to trade instantly against a limit order, or to post a limit order. She cannot monitor her limit order continuously, in that she can only revise or cancel her order at her next contact time with the market. It is shown that, if price jumps are frequent enough and if the tail of the jump size is fat enough (representing a high risk of being picked off), then a seller prefers to trade instantly when she contacts the market, rather than post a limit order. Because the market price is rising on average, a buyer also prefers to purchase instantly at a low market price, rather than later with a limit order. Therefore, in the equilibrium of Addendum I, the limit order book is empty, liquidity is provided only by marketmakers, and the asset allocation is the same as that of Theorem 2.

4.2 Borrowing-Constrained Marketmakers

The implementation result of Theorem 2 relies on the assumption that the time-zero capital \(a(0)\) is sufficiently large. This ensures that, in equilibrium, a marketmaker’s borrowing constraint (45) never binds. There is, however, much anecdotal evidence suggesting that, during the October 1987 crash, specialists’ and marketmakers’ borrowing constraints were binding. Some market commentators have suggested that insufficient capital might have amplified the disruptions (see, among others, Brady (1988) and Bernanke (1990)). This subsection describes an amplification mechanism associated with insufficient capital and binding bor-

\(^{20}\)The risk of being picked off is often described as a “winner’s curse” problem. Indeed, a limit order to sell is more likely to be executed when the fundamental value of the asset goes up, and vice versa for a limit order to buy.
rowing constraints. Specifically, the following proposition shows that if marketmakers are borrowing constrained during the crash and if their time-zero capital is small enough, then they do not have enough purchasing power to absorb the selling pressure, and therefore fail to provide the optimal amount of liquidity.

**Proposition 6 (Equilibrium with small capital).** There exists \( a_0^* \leq \pi_0^* \) such that:

(i) If marketmakers’ aggregate capital is \( a(0) \in [0, a_0^*) \), there exists an equilibrium whose allocation is a buffer allocation with maximum inventory position \( m \in [0, m^*) \).

(ii) If marketmakers’ aggregate capital is \( a(0) \in [a_0^*, \pi_0^*) \), there exists an equilibrium whose allocation is a buffer allocation with maximum inventory position \( m^* \).

If \( t_1^* > 0 \), then \( a_0^* = \pi_0^* \) and the interval \( [a_0^*, \pi_0^*) \) is empty. Lastly, in all of the above, the equilibrium price path has a strictly positive jump at the time \( t_m \) such that \( I(t_m) = m \) (that is, \( t_m = \psi(m) \) for the function \( \psi(\cdot) \) of Proposition 1).

The equilibrium price and allocation are shown in Figures 7 and 8. The price jumps up at time \( t_m \). It grows at a low rate \( \dot{p}(t)/p(t) < r \) for \( t \in [0, t_1) \), at a high rate \( \dot{p}(t)/p(t) = r \) for \( t \in (t_1, t_m) \cup (t_m, t_2) \), and at a zero rate after \( t_2 \).

The price jump implies that a marketmaker makes positive profit. For example, a marketmaker can buy assets the last instant before the jump at a low price \( p(t_m^-) \) and re-sell these assets the next instant after the jump at the strictly higher price \( p(t_m^+) \). An optimal trading strategy maximizes the profit that a marketmaker extracts from the price jump, as follows: i) a marketmaker invests all of her capital at the risk-free rate during \( t \in [0, t_1) \) in order to increase her buying power, ii) spends all her capital in order to buy assets before the jump, during \( t \in (t_1, t_m) \), iii) re-sells all her assets after the jump, during \( (t_m, t_2) \). A marketmaker does not hold any assets during \( t \in [0, t_1) \) and \( t \in (t_2, \infty) \) because the price grows at a rate strictly less than \( r \). Because the price grows at rate \( r \) during \( (t_1, t_m) \) and \( (t_m, t_2) \), a marketmaker is indifferent regarding the timing of her purchases and sales, as long as all assets are purchased during \( (t_1, t_m) \), sold during \( (t_m, t_2) \), and all capital is used up at time \( t_m \).

The price jump at time \( t_m \) seems to suggest the following arbitrage: a utility-maximizing marketmaker would buy more assets shortly before \( t_m \) and sell them shortly after. This does not, in fact, truly represent an arbitrage, because a marketmaker runs out of capital precisely at the jump time \( t_m \), so she cannot purchase more assets.

27
Perhaps the most surprising result is that, if $t_1^* = 0$, then there is a non-empty interval $[a_0^*, \bar{a}_0^*]$ of time-zero capital such that the price has a positive jump and marketmakers accumulate the optimal amount $m^*$ of inventories. If $t_1^* > 0$, then $a_0^* = \bar{a}_0^*$ and the interval is empty. Intuitively, if the interval were not empty, then a small increase in time-zero capital combined with a positive price jump would give marketmakers incentive to provide more liquidity. Hence, they would start buying assets at some time $t_1 < t_1^*$ and would end up accumulating more inventories than $m^*$, which would be a contradiction. If $t_1^* = 0$, this reasoning does not apply: indeed, an increase in time-zero capital cannot increase inventory accumulation because marketmakers cannot start accumulating inventories earlier than time zero.
5 Policy Implications

This section discusses some policy implications of this model of optimal liquidity provision.

Marketmaking capital

The model suggests that, with perfect capital markets, competitive marketmakers would have enough incentive to raise sufficient capital. The intuition is that marketmakers will raise capital until their net profit is equal to zero, which precisely occurs when they provide optimal liquidity. For example, suppose that, at $t = 0$, wealthless marketmakers can borrow capital instantly on a competitive capital market. Then, for $t > 0$, the economic environment remains the one described in the present paper. If, at $t = 0$, a marketmaker borrows a quantity $a > 0$, then she has to repay $ae^{rT}$ at some time $T \geq t^*$. One can show that, with an optimal trading strategy described at the end of Section 4.2, the net present value of her profit is

$$\left(\frac{p(t_+^m)}{p(t_-^m)} - 1\right)a, \quad (48)$$

where the jump-size $(p(t_+^m)/p(t_-^m) - 1)$ depends implicitly on the time-zero aggregate market-making capital.\(^{21}\) As long as the jump size is strictly positive, a marketmaker wants to borrow an infinite amount of capital. Therefore, in a capital-market equilibrium, a marketmaker’s net profit \((48)\) must be zero, implying that $p(t_+^m)/p(t_-^m) = 1$ and $m = m^*$. This means that marketmakers borrow a sufficiently large amount of capital and provide the socially optimal amount of liquidity.

Lending capital to marketmakers, however, might be costly because of capital-market imperfections associated for example with moral hazard or adverse selection problems. In order to compensate for such lending costs, the net return $(p(t_+^m)/p(t_-^m) - 1)$ on marketmaking capital must be greater than zero. This would imply that, in an equilibrium, marketmakers do not raise sufficient capital. As a result, subsidizing loans to marketmakers might improve welfare.\(^{22}\) The model supports the recommendation of subsidizing loans to marketmakers.

\(^{21}\)The profit \((48)\) is not discounted by $e^{-rT}$ because a marketmaker can invest her capital at the risk free rate.

\(^{22}\)Addendum II provides an explicit model of marketmakers’ borrowing limits based on moral hazard. A different model of limited access to capital is due to Shleifer and Vishny (1997). They show that capital constraints might be tighter when prices drop, due to a backward-looking, performance-based rule for allocating capital to arbitrage funds.
During disruptions, some policy actions can be interpreted as bank-loan subsidization. For instance, during the October 1987 crash, the Federal Reserve lowered the fund rate, while encouraging commercial banks to lend generously to security dealers (Wigmore (1998)).

Capital requirements for imperfectly competitive marketmakers

In an article published by The Financial Times, Maurice “Hank” Greenberg, chairman of the American International Group (AIG), severely criticizes the seven specialist firms of NYSE which handle the trading of more than 2,800 stocks. In particular, he argues that these firms do not maintain sufficient capital and he proposes to raise their minimum regulatory capital requirements. Suppose for simplicity that there are two marketmakers who choose simultaneously their time-zero capital, and who compete in price for the order flow. Assume that competition in price implies the perfectly competitive outcome, meaning that the equilibrium during the crash with two marketmakers is the same as with a continuum of marketmakers. The choice of time-zero capital, however, would be different than with a continuum of marketmakers. A marketmaker would recognize that committing less capital to marketmaking would increase her net return during the crash. Hence, she would have incentive to maintain a small level of capital so as to make strictly positive profit, meaning that the aggregate marketmaking capital would turn out to be smaller than optimal (an intuition analogous to Kreps and Scheinkman (1983)). One assumes that, in a model which incorporates this intuition, the setting of a minimum regulatory capital requirement could improve welfare. Such a result would support Greenberg’s claim that specialists firms are undercapitalized and should face tighter capital requirements.

Price continuity

It is often argued that marketmakers should provide liquidity in order to maintain price continuity and to smooth asset price movements. The present paper studies liquidity provision in terms of the Pareto criterion rather than in terms of some price smoothing objective. The

---

23 The article “Shake up the NYSE Specialist System or Drop it” was published by The Financial Times on October 10th, 2003.

24 For instance, Investor Relations, an advertising document for the specialist firm Fleet Meehan Specialist, argues that specialists “use their capital to fill temporary gaps in supply and demand. This can actually help to reduce short-term volatility by cushioning the intra-day price movements.”
results are evidence that Pareto optimality is consistent with a discrete price decline at the time of the crash. This suggests that requiring marketmakers to maintain price continuity at the time of the crash may result in a welfare loss.

A comparative static exercise suggests, however, that liquidity provision promotes some degree of price continuity. Namely, in an economy with no capital at time zero \((a(0) = 0)\), no liquidity is provided and the price jumps up at time \(t_s > 0\). In an economy with large time-zero capital, however, the price path is continuous at each time \(t > 0\).

*Marketmakers as Buyers of Last Resort*

A commonly held view is that marketmakers should not merely provide liquidity, they should also provide it promptly. In contrast with that view, the present model illustrates that prompt action is not necessarily consistent with efficiency. Namely, it is not always optimal that marketmakers start providing liquidity immediately at the time of the crash, when the selling pressure is strongest. For example, if the initial preference shock is very persistent, then marketmakers who buy asset immediately end up holding assets for a very long time. This cannot be efficient given that marketmakers are not the final holders of the asset. This suggests that requiring marketmakers to always buy assets immediately at the time of the crash can result in a welfare loss.

6 Conclusion

This paper studies the optimal liquidity provision of marketmakers during financial disruptions. The first main result is that competitive marketmakers will provide the optimal amount of liquidity, provided they maintain sufficient capital at the time of the crash. If capital-market imperfections prevent marketmakers from raising sufficient capital before the crash, transferring purchasing power to marketmakers during the crash might improve welfare. The second main result is that the competitive equilibrium has features which are traditionally viewed as symptomatic of poor liquidity provision but are in fact consistent with efficiency. Namely, there is a discrete price decline at the time of the crash and marketmakers do not always start buying assets immediately when the selling pressure is strongest.
A Proof of Proposition 1

In all what follows, some buffer allocation \((t_1, t_2)\) is fixed.

**Hump Shape.** One first shows that \(I(t)\) is hump-shaped and that, given the first breaking time \(t_1\), the second breaking time \(t_2\) is uniquely characterized. For \(t \in [t_1, t_2)\), the inventory position \(I(t)\) evolves according to \(\dot{I}(t) = u(t) - u_h(t) = \rho(\mu(t) - \mu_h(t))\). With equation (3), this ODE can be written
\[
\dot{I}(t) = -\rho I(t) + \rho(s - \mu_h(t)).
\] (49)
Together with the initial condition \(I(t_1) = 0\), this implies that, for \(t \in [t_1, t_2]\), \(I(t) = H(t_1, t)\), where
\[
H(t_1, t) = \rho \int_{t_1}^{t} (s - \mu_h(z)) e^{\rho(z-t)} dz
\] (50)
One has \(\partial H/\partial t = -\rho H + \rho(s - \mu_h(t))\), implying that \(\partial H/\partial t(t_1, t_1) = \rho(s - \mu_h(t_1)) \geq 0\) and \(\partial H/\partial t(t_1, t_s) = -\rho^2 \int_{t_1}^{t_s} (s - \mu_h(z)) e^{\rho(z-t_s)} dz \leq 0\). Therefore, there exists \(t_m \in [t_1, t_s]\) such that \(\partial H/\partial t(t_1, t_m) = 0\). Moreover, \(\partial H/\partial t = 0\) implies that \(\partial^2 H/\partial t^2 = -\rho \dot{\mu}_h(t) - \rho \partial H/\partial t = -\rho \dot{\mu}_h(t) < 0\). This implies that \(t_m\) is unique, that \(H(t_1, t)\) is strictly increasing for \(t \in [t_1, t_m)\), and strictly decreasing for \(t \in (t_m, \infty)\). Now, because \(\mu_h(t) > s\) for \(t\) large enough, it follows from (50) that \(H(t_1, t)\) is negative for \(t\) large enough. Therefore, given some \(t_1 \in [0, t_s]\), there exists a unique \(t_2 \in [t_s, +\infty)\) such that \(H(t_1, t_2) = 0\).

Writing \(\{t_1, t_m, t_2\}\) as a function the the maximum inventory position. The maximum inventory position of Proposition 1 is defined as \(m \equiv I(t_m)\). One let \(t_m \equiv \psi(m)\), for some function \(\psi(\cdot)\) which can be written in closed form by substituting \(\dot{I}(t_m) = 0\) and \(I(t_m) = m\) in (49):
\[
\psi(m) = -\frac{1}{\gamma} \log \left(1 - \frac{s - m}{y}\right).
\] (51)
Now, solving (49) with the initial condition \(I(t_m) = m\), one finds
\[
I(t) = me^{-\rho(t-t_m)} + (s - y)(1 - e^{-\rho(t-t_m)}) + pye^{-\gamma t_m} e^{-\rho(t-t_m)} \int_{t_m}^{t} e^{(\rho-\gamma)u} du.
\] (52)
Replacing (51) into (52), and making some algebraic manipulations, show that \(I(t) = 0\) if and only if \(t = t_m + z\), for some \(z\) solution of \(G(m, z) = 1\), where
\[
G(m, z) = \left(1 + \frac{m}{y-s}\right) e^{-\rho z} \left[1 + \rho \int_{0}^{z} e^{(\rho-\gamma)u} du\right].
\] (53)
Let's define, for \(x \in [0, +\infty)\), the two functions \(g_i(m, x) = G(m, (-1)^i \sqrt{x})\), \(i \in \{1, 2\}\). For \(x > 0\), the partial derivatives of \(g_i\) with respect to \(x\) is
\[
\frac{\partial g_i}{\partial x} = \left(1 + \frac{m}{y-s}\right) \frac{(-1)^i \rho e^{-\rho(1)^i \sqrt{x}}}{2\sqrt{x}} \times \left[-1 - \rho \int_{0}^{(-1)^i \sqrt{x}} e^{(\rho-\gamma)u} du + e^{(\rho-\gamma)(-1)^i \sqrt{x}}\right]
\] (54)
It can be extended by continuity at \( x = 0 \) as

\[
\frac{\partial g_i}{\partial x}(m, 0) = -\left(1 + \frac{m}{y-s}\right)\frac{\rho_\gamma}{2}.
\] (55)

The term in bracket in (54) is zero at \( x = 0 \), and is easily shown to be strictly increasing (decreasing) for \( i = 1 \) (\( i = 2 \)). Together with (55), this shows that \( g_i(m, \cdot) \) is strictly decreasing over \([0, +\infty)\). Moreover, for \( m = 0, \ g_i(0, 0) = 1 \). For \( m > 0, g_i(m, 0) > 1, g_1(m, x) \to -\infty \) and \( g_2(m, x) \to 0 \) when \( x \to +\infty \). This implies that, for any \( m \geq 0 \), there exists only one solution \( x_i = \Phi_i(m) \) of \( g_i(m, x) = 1 \). An application of the Implicit Function Theorem (see Taylor and Mann (1983), Chapter 12) shows that the function \( \Phi_i(\cdot) \) is strictly increasing and continuously differentiable, and satisfies \( \Phi_i(0) = 0, \Phi_i'(0) = 2/(\rho\gamma(y - s)) \). Clearly, \( G(m, z) = 0 \) if and only if \( z \in \{\phi_1(m), \phi_2(m)\} \), with \( \phi_i(m) = (-1)^i\sqrt{\Phi_i(m)} \). Lastly, the restriction \( t_1 \geq 0 \) defines the domain of the functions \( \psi(\cdot), \phi_1(\cdot), \) and \( \phi_2(\cdot) \). Namely, \( t_1 \geq 0 \) if and only if

\[
\psi(m) - \phi_1(m) \geq 0.
\] (56)

The left-hand side of (56) is strictly decreasing, is strictly positive for \( m = 0 \) and strictly negative for \( m = s \). Hence, there exists a unique \( \bar{m} \) such that \( \psi(\bar{m}) - \phi_1(\bar{m}) = 0 \). By construction, the maximum inventory of a buffer allocation is less than \( \bar{m} \).

## B  Socially Optimal Allocations

This appendix solves for socially optimal allocations. In order to prove the various results of Section 3 and 4, it is convenient to assume that, in addition to the trading technology and the short-selling constraint, the planner is also constrained by an inventory bound

\[
I(t) \leq M,
\] (57)

for some \( M \in [0, +\infty] \).

### B.1 First-Order Sufficient Conditions

The current-value Lagrangian and the first-order conditions are the one of Subsection 3.3, with an additional multiplier \( \eta_M(t) \) for the inventory bound (57). It is important to note that, because of the inventory bound, the multiplier \( \lambda_I(t) \) can also jump up, with the restrictions \( \lambda_I(t^+) - \lambda_I(t^-) \geq 0 \) if \( I(t) = M \). In what follows, it is convenient to eliminate \( \lambda_h(t) \) and \( \lambda_(t) \) from the first-order conditions using (28) and (29). One obtains the reduced system

\[
\begin{align*}
rw(t) &= \delta - 1 - \gamma_u(w_h(t) + w(t)) - \rho w(t) \\
+ \eta_I(t) - \eta_M(t) + \hat{\omega}(t) \\
rw_h(t) &= 1 - \gamma_d(w_h(t) + w(t)) - \rho w(t) \\
- \eta_I(t) + \eta_M(t) + \hat{w}(t) \\
r\lambda_I(t) &= \eta_I(t) - \eta_M(t) + \hat{\lambda}_I(t),
\end{align*}
\] (58)

(59)

(60)

with the jump conditions

\[
\begin{align*}
\lambda_I(t^+) - \lambda_I(t^-) \leq 0 & \text{ if } I(t) = 0 \\
\lambda_I(t^+) - \lambda_I(t^-) \geq 0 & \text{ if } I(t) = M \\
\lambda_I(t^+) - \lambda_I(t^-) &= w_t(t^+) - w_t(t^-) = -w_h(t^+) + w_h(t^-),
\end{align*}
\] (61)

(62)

(63)
and the transversality conditions

\[
\lim_{t \to +\infty} e^{-rt}\lambda_I(t) = \lim_{t \to +\infty} e^{-rt}w_\ell(t) = \lim_{t \to +\infty} e^{-rt}w_h(t) = 0. \tag{64}
\]

As before, the positivity restrictions and complementary slackness conditions are

\[
\begin{align*}
\eta_I(t) &\geq 0 \quad \text{and} \quad \eta_I(t)I(t) = 0 \quad \text{(65)} \\
\eta_M(t) &\geq 0 \quad \text{and} \quad \eta_M(t)(M - I(t)) \quad \text{(66)} \\
w_\ell(t) &\geq 0 \quad \text{and} \quad w_\ell(t)(\rho \mu_{\ell\ell}(t) - u_\ell(t)) = 0 \quad \text{(67)} \\
w_h(t) &\geq 0 \quad \text{and} \quad w_h(t)(\rho \mu_{h\ell}(t) - u_h(t)) = 0. \quad \text{(68)}
\end{align*}
\]

Because the optimization problem is linear, there is no sign restrictions on the multiplier \(\lambda_h(t)\) and \(\lambda_\ell(t)\) (see Section 3 in Part II of Kamien and Schwartz (1991)). As a result, the reduced system (58)-(68) of first-order condition is equivalent to the original system of first-order sufficient conditions.

### B.2 Multipliers for Buffer Allocations

Consider some feasible buffer allocation with breaking times \((t_1, t_2)\) and a maximum inventory position \(m \in [0, \min\{\bar{m}, M\}]\) reached at time \(t_m\). This paragraph first constructs a collection \((w_\ell(t), w_h(t), \lambda_\ell(t), \eta_I(t), \eta_M(t))\) of multipliers solving equations (58)-(68), but ignoring some of the positivity restrictions. These restrictions are imposed afterwards, when discussing the optimality of this allocation. First, one guesses that \(\eta_M(t) = 0\) for all \(t \geq 0\). Second, summing equations (58) and (59), and using the transversality condition (64) shows that

\[
w_\ell(t) + w_h(t) = \frac{\delta}{r + \rho + \gamma}, \tag{69}
\]

for all \(t \geq 0\). Then, one guesses that there are no jumps at \(t_1\) and \(t_2\). With (63), this shows that \(\lambda_\ell(t_1^+) - \lambda_\ell(t_1^-) = w_\ell(t_1^+) - w_\ell(t_1^-) = -w_h(t_1^-) + w_h(t_1^-) = 0\), for \(i \in \{1, 2\}\). Now, one can solve for the multipliers, going backwards in time.

**Time Interval** \(t \in [t_2, +\infty)\). Complementary slackness (68) implies that \(w_h(t_2) = 0\). With (69), this shows that \(w_\ell(t) = \frac{\delta}{r + \rho + \gamma}\). With (59) and (69), this also implies that \(\eta_I(t) = -\gamma_d \delta / (r + \rho + \gamma)\). Lastly, (60) and (64) show that \(r \lambda_I(t) = 1 - \delta \gamma_d / (r + \rho + \gamma)\).

**Time Interval** \(t \in [t_m, t_2)\). First, because \(I(t) > 0\), the complementary slackness condition (65) implies that \(\eta_I(t) = 0\). Then, one solves the ODE (59) with the terminal condition \(w_h(t_2) = 0\), and one finds that

\[
w_h(t) = \frac{1}{r + \rho} \left(1 - \delta \frac{\gamma_d}{r + \rho + \gamma}\right) (1 - e^{(r + \rho)(t-t_2)}), \tag{70}
\]

for \(t \in [t_m, t_2)\). With (69), \(w_\ell(t) = \delta / (r + \rho + \gamma) - w_h(t)\). Similarly, one can solve the ODE (60) with the terminal condition \(r \lambda_I(t_2) = 1 - \gamma_d \delta / (r + \rho + \gamma)\), finding that

\[
r \lambda_I(t) = \left(1 - \delta \frac{\gamma_d}{r + \rho + \gamma}\right) e^{r(t-t_2)}. \tag{71}
\]
Time Interval \( t \in [t_1, t_m) \). In this time interval, \( \eta_I(t) = 0 \). One needs to consider two cases.

**Case 1:** \( m < \bar{m} \). Complementary slackness at \( t = t_1 \) shows that \( w(t_1) = 0 \), implying that \( w(t_1) = \delta/(r + \rho + \gamma) \). With this and (59), one finds

\[
w_h(t) = \frac{1}{r + \rho} \left( 1 - \delta \frac{\gamma d}{r + \rho + \gamma} - \left( 1 - \delta \frac{r + \rho + \gamma d}{r + \rho + \gamma} \right) e^{(r+\rho)(t-t_1)} \right),
\]

for \( t \in [t_1, t_m) \). Given (70) and (72), the multiplier \( w_h(t) \) is not necessarily continuous at time \( t_m \). The size \( w_h(t_m) - w_h(t_m^-) \) of the jump can be written as some function \( b(\cdot) \) of the maximum inventory level \( m \), where

\[
b(m) = \frac{1}{r + \rho} \left[ \left( 1 - \delta \frac{\gamma d}{r + \rho + \gamma} \right) e^{-(r+\rho)\phi_2(m)} - \left( 1 - \delta \frac{r + \rho + \gamma d}{r + \rho + \gamma} \right) e^{(r+\rho)\phi_1(m)} \right].
\]

Equation (63) implies that \( \lambda_f(t_m^-) - \lambda_f(t_m^+ - b(m) \right) e^{r(t-t_m)}.
\]

**Case 2:** \( m = \bar{m} \). Then, by construction of \( \bar{m} \), \( t_1 = 0 \). If \( b(\bar{m}) < 0 \), the multipliers are constructed as in Case 1. If, on the other hand, \( b(\bar{m}) \geq 0 \), then one constructs a set of multipliers by solving the ODEs (59) and (60) with terminal conditions \( w_h(t_m^-) = w_h(t_m^+) + (1-\alpha)b(\bar{m}) \) and \( \lambda_f(t_m^-) = \lambda_f(t_m^+) - (1-\alpha)b(\bar{m}) \), for all \( \alpha \in [0, 1] \). That \( b(\bar{m}) \geq 0 \) implies that, for all \( \alpha \in [0, 1] \), \( w_h(t_1) = w_h(0) \in \left[0, \delta/(r + \rho + \gamma)\right] \). By construction, the \( \alpha = 1 \) multipliers do not jump at \( t = t_m \).

Time Interval \( t \in [0, t_1] \), \( m < \bar{m} \) Complementary slackness shows that \( w(t) = 0 \), implying that \( w(t) = \delta/(r + \rho + \gamma) \). With equation (58), this also implies that \( \eta_I(t) = 1-\delta(r+\rho+\gamma d)/(r+\rho+\gamma) \geq 0 \). Then

\[
r\lambda_f(t) = -\eta_I(t) + (\lambda_f(t_1^-) + \eta_I(t))e^{r(t-t_1)}.
\]

**B.3 Proof of Proposition 2**

Let’s consider the multipliers associated with the no-inventory allocation \( M \), as constructed in Section B.2. It is shown that the multiplier \( \lambda_f(\cdot) \) jumps at the crossing time \( t_s \), with \( \lambda_f(t_s^+) - \lambda_f(t_s^-) = \delta/(r + \rho + \gamma) \). This positive jump reflects the benefit of accumulating inventories near the crossing time. These multipliers can be used to compare the no-inventory allocation with other allocation. In particular, for any buffer allocation \((\mu^m, I^m, u^m)\),

\[
W(m) - W(0) = - \int_0^{t_s} e^{-rt} \left( \rho \mu^m_h(t) - u^m_h(t) \right) dt - \int_0^{t_s} e^{-rt} \left( \rho \mu^m_d(t) - u^m_d(t) \right) dt - \int_0^{t_s} e^{-rt} \eta_I(t) I^m(t) dt + (\lambda_f(t_s^+) - \lambda_f(t_s^-)) I^m(t_s) e^{-r t_s}.
\]
Formula (76) follows from the standard comparison argument of Optimal Control (see, for example, Section 3 of Part II in Kamien and Schwartz (1991)), in the special case of a linear objective and linear constraints. The first two terms in equation (76) are zero because, for all \( t \leq t_s \), \( w^m_h(t) = \rho \mu^m_h(t) \) and \( w^\ell(t) = 0 \), and, for all \( t \geq t_s \), \( w^m_t(t) = \rho \mu^m_t(t) \) and \( w_h(t) = 0 \). The third term can be bounded as follows

\[
0 \leq \int_{\psi(m) - \phi_1(m)}^{\psi(m) + \phi_2(m)} e^{-rt} \eta(t) I^m(t) dt \leq \left( 1 - \frac{\gamma_d - \delta}{r + \rho + \gamma} \right) m (\phi_2(m) + \phi_1(m)).
\]

Since \( \lim_{m \to 0^+} \phi_i(m) = 0 \), for \( i \in \{1, 2\} \), this implies that, as \( m \to 0^+ \),

\[
\frac{1}{m} \int_{\psi(m) - \phi_1(m)}^{\psi(m) + \phi_2(m)} e^{-rt} \eta(t) I^m(t) dt \to 0.
\]

(77)

The fourth and last term of (76) is

\[
(\lambda_I(t^+_s) - \lambda_I(t^-_s)) e^{-r t_s} I^m(t_s).
\]

(78)

Equation (52) shows that \( I^m(t_s) = 0 \) at \( m = 0 \). Using the facts that \( t_s - t_m = \psi(0) - \psi(m) \), and that \( y e^{-\gamma t m} = y - s + m \), differentiating (52) with respect to \( m \) shows that the derivative of \( I^m(t_s) \) at \( m = 0 \) is equal to one. Therefore

\[
\frac{1}{m} I^m(t_s) \to 1,
\]

(79)

as \( m \) goes to zero, establishing Proposition 2.

### B.4 Proof of Theorem 1

This paragraph verifies that some buffer allocation is constrained-optimal with inventory bound \( M \). First, if some buffer allocation is constrained-optimal, it must satisfy the jump condition (62), meaning that \( b(m) \geq 0 \) and \( b(m)(M - m) = 0 \). In particular, if there is no inventory constraint, then the jump must be zero. One defines the maximum \( m \) such that the jump \( b(m) \) is positive:

\[
m^* = \sup\{m \in [0, \bar{m}] : b(m) \geq 0\}.
\]

(80)

If \( m^* < \bar{m} \), then \( b(m^*) = 0 \). If \( m^* = \bar{m} \), then \( b(m^*) > 0 \). Furthermore, since \( b(\cdot) \) is decreasing, \( b(m) \geq 0 \) for all \( m \leq m^* \).

**Proposition 7.** For all \( M \in [0, \infty] \), the buffer allocation with maximum inventory position \( m = \min\{m^*, M\} \) is socially optimal with inventory bound \( M \).

In order to prove this proposition, let’s consider this allocation and its associated multipliers, constructed as in Appendix B.4. (If \( m^* = \bar{m} \), one picks the multipliers which do not jump at time \( t_m \)). Two optimality conditions remain to be verified: the jump conditions (62) and the positivity restrictions in (67) and (68). Because \( m \leq m^* \), the jump condition (62) is satisfied. Also, because \( w_h(t) \) is a decreasing function of time, \( w_h(0) \in [0, \delta/(r + \rho + \gamma)] \) and \( w_h(t_2) = 0 \), it follows that, at each time, \( w_h(t) \in [0, \delta/(r + \rho + \gamma)] \), and therefore that \( w_\ell(t) \geq 0 \).
The inventory-accumulation period $\Delta^*$ and the breaking times $(t_1^*, t_2^*)$ of Theorem 1 are found as follows. First, $\Delta^* = \phi_1(m^*) + \phi_2(m^*)$. Then, simple algebraic manipulations show that $b(m) \geq 0$ if and only if
\[
e^{(r+\rho)(\phi_1(m) + \phi_2(m))} \leq 1 + \frac{\delta(r + \rho)}{\gamma_u + (1 - \delta)(r + \rho + \gamma_d)}.
\]
(81)
If $m^* < \bar{m}$, then (81) holds with equality at $m^*$, and if $m^* = \bar{m}$, it holds with inequality. This is equivalent to the formula of Theorem 1. Then, given $\Delta^*$, the first breaking time $t_1^*$ is a solution of $H(t_1^*, t_1^* + \Delta^*) = 0$. Direct integration of (50) shows that
\[
H(t_1, t) = (s - y) \left(1 - e^{-\rho(t-t_1)}\right) + \rho y e^{-\gamma t_1} \frac{e^{-\gamma(t-t_1)} - e^{-\rho(t-t_1)}}{\rho - \gamma},
\]
where we let $(e^{-\gamma x} - e^{-\rho x})/(\rho - \gamma) = x$ if $\rho = \gamma$. Simple algebraic manipulation of (82) give the analytical solution of Theorem 1.

B.5 Proof of Proposition 3

The optimal allocation is denoted $(\mu^*(t), I^*(t), u^*(t))$. The multipliers associated with this allocation are denoted $(w_h(t), w_e(t), \lambda_I(t), \eta_I(t))$. These are continuous. Let’s consider another allocation $(\mu(t), I(t), u(t))$ which achieves the optimum. The standard comparison argument of Optimal Control (see Section 3 of Part II in Kamien and Schwartz (1991)) implies that
\[
\int_0^{+\infty} e^{-rt} (w_h(t)(\rho \mu_h(t) - u_h(t)) + w_e(t)(\rho \mu_e(t) - u_e(t)) + \eta_I(t)I(t)) dt = 0.
\]
Each term in the integrand is positive, and therefore is equal to zero, almost everywhere. Then, because $\eta_I(t)I(t) = 0$, it must be that $u_h(t) = u_e(t)$ almost everywhere in $[0, t_1] \cup [t_2, +\infty)$. In $[0, t_1]$, $w_h(t) > 0$ and therefore $u_h(t) = \rho \mu_h(t) = u_e(t)$ almost everywhere. In $[t_2, +\infty)$, $w_e(t) > 0$ and therefore $u_e(t) = \rho \mu_e(t) = u_h(t)$ almost everywhere. Lastly, in $[t_1, t_2]$, both $w_h(t) > 0$ and $w_e(t) > 0$, implying that $u_h(t) = \rho \mu_h(t)$, and $u_e(t) = \rho \mu_e(t)$, almost everywhere. Therefore, the allocation $(\mu(t), I(t), u(t))$ is equal to $(\mu^*(t), I^*(t), u^*(t))$, almost everywhere. This proves Proposition 3.

B.6 Proof of Proposition 4

From Theorem 1, the length of the inventory accumulation period is $\Delta^* = \min\{D(x), \bar{\Delta}\}$, where
\[
D(x) = \frac{1}{r + \rho} \log \left(1 + \frac{\delta(r + \rho)}{\gamma_u + (1 - \delta)(r + \rho + \gamma_d)}\right).
\]
(83)
The function $D(\cdot)$ is clearly increasing in $\delta$, decreasing in $\gamma_d$ and $\gamma_u$. It remains to show that it is also decreasing in $r$ and $\rho$. Holding $(\delta, \gamma_u, \gamma_d)$ fixed, and letting $z \equiv r + \rho$, one let $D(x) = \frac{1}{z} \log(1 + \alpha(z)) \equiv \eta(z)$ where $\alpha(z) = z/(a + bz)$ for some $(a, b) \in \mathbb{R}^2_+$. The first derivative of $\eta(\cdot)$ is
\[
\frac{d\eta}{dz} = -\frac{1}{z^2} \left(\log(1 + \alpha(z)) - \frac{z\alpha'(z)}{1 + \alpha(z)}\right) \equiv -\frac{1}{z^2} \beta(z),
\]
(84)
where \( \alpha'(z) \) denotes \( d\alpha/dz \). In turn,

\[
\frac{d\beta}{dz} = -z \frac{d}{dz} \left( \frac{\alpha'(z)}{1 + \alpha(z)} \right).
\]

Since \( \alpha'(z)/(1 + \alpha(z)) = a/(a + bz)/(a + (1 + b)z) \) is decreasing for \( z \geq 0 \), it follows that that \( d\beta/dz \geq 0 \). Since \( \beta(0) = 0 \), it follows that \( \beta(z) \geq 0 \). Therefore, (84) implies that \( d\eta/dz \leq 0 \). This establishes that \( D( \cdot ) \) is decreasing in \( r \) and \( \rho \).

### B.7 Proof of Proposition 5

The maximum inventory position is reached at time \( t^*_m \). Theorem 1 implies that \( \Delta^* = t^*_2 - t^*_1 \to 0 \), as \( \rho \to +\infty \). Since \( t^*_1 < t^*_m < t^*_s < t^*_2 \), it also implies that \( t^*_m \to t^*_s \), as \( \rho \to +\infty \). Using the function \( \psi(\cdot) \) of (51), one can write \( t^*_m = -1/\gamma \log (1 + (s - m^*)/y) \). Since this equation does not depend on \( \rho \), it follows that \( m^* \to 0 \) as \( \rho \to +\infty \).

### C Market Equilibria

This appendix provides a precise definition of a competitive equilibrium, and proves Theorem 2 and Proposition 6.

#### C.1 Definition of a Competitive Equilibrium

First, investors’ continuation utilities solve the ODE

\[
\begin{align*}
\rho V_{\ell n} &= \gamma_u (V_{hn} - V_{\ell n}) + \dot{V}_{\ell n} \tag{86} \\
\rho V_{lo} &= 1 - \delta + \gamma_u (V_{ho} - V_{lo}) + \rho (V_{\ell n} - V_{lo} + p) + \dot{V}_{lo} \tag{87} \\
\rho V_{hn} &= \gamma_d (V_{\ell n} - V_{hn}) + \rho (V_{ho} - V_{hn} - p) + \dot{V}_{hn} \tag{88} \\
\rho V_{ho} &= 1 + \gamma_d (V_{lo} - V_{ho}) + \dot{V}_{ho}, \tag{89}
\end{align*}
\]

where the price \( p \) and the continuation utilities \( V_\sigma \) are implicitly function of time and \( \dot{V}_\sigma \equiv dV_\sigma(t)/dt \). Hence, the reservation value \( \Delta V_\ell \) and \( \Delta V_h \) of a seller and of a buyer solve

\[
\begin{align*}
\rho \Delta V_\ell &= 1 - \delta + \gamma_u (\Delta V_h - \Delta V_\ell) + \rho (p - \Delta V_\ell) + \Delta \dot{V}_\ell \tag{90} \\
\rho \Delta V_h &= 1 + \gamma_d (\Delta V_\ell - \Delta V_h) - \rho (\Delta V_h - p) + \Delta \dot{V}_h. \tag{91}
\end{align*}
\]

Lastly, in order to complete the standard optimality verification argument, I impose the transversality conditions

\[
\lim_{t \to +\infty} \Delta V_j(t)e^{-rt} = 0, \tag{92}
\]

for \( j \in \{\ell, h\} \). Conversely, given the reservation values, one finds the continuation utilities by solving the ODE (86) and (88) for \( V_{\ell n} \) and \( V_{hn} \), and by letting \( V_{jo} = V_{jn} + \Delta V_j \) for \( j \in \{h, \ell\} \). Namely, subtracting (86) from (88), integrating, and assuming transversality, one finds that

\[
V_{hn}(t) - V_{\ell n}(t) = \int_t^{+\infty} e^{-(r+\gamma)(z-t)} \rho (\Delta V_h(z) - p(z)) \, dz \tag{93}
\]
and, replacing (93) in (86), that
\[
V_{\ell n}(t) = \int_{t}^{+\infty} e^{-r(z-t)} \gamma_u (V_{hn}(z) - V_{\ell n}(z)) dz.
\]

**Definition 4 (Competitive Equilibrium).** A Competitive Equilibrium is a feasible allocation \((\mu(t), I(t), u(t))\), a price \(p(t)\), a collection \((\Delta V_\ell(t), \Delta V_h(t))\) of reservation values, a consumption stream \(c(t)\), and a bank account position \(a(t)\) such that:

(i) given the price \(p(t)\), \((I(t), u(t), c(t), a(t))\) solves the marketmaker’s problem, and

(ii) given the price \(p(t)\), the reservation values \((\Delta V_\ell(t), \Delta V_h(t))\) solve equations (90)-(92) and satisfy, at each time,

\[
\begin{align*}
 p(t) - \Delta V_\ell(t) &\geq 0 & (94) \\
 \Delta V_h(t) - p(t) &\geq 0 & (95) \\
 (p(t) - \Delta V_\ell(t))(\rho \mu_{\ell o}(t) - u_\ell(t)) &= 0 & (96) \\
 (\Delta V_h(t) - p(t))(\rho \mu_{hn}(t) - u_h(t)) &= 0. & (97)
\end{align*}
\]

Equations (94) through (97) verify the optimality of investors’ policies. For instance, equation (94) means that the net utility of selling is positive, which verifies that a seller \(\ell o\) finds it weakly optimal to sell.\(^{25}\) Equation (96), on the other hand, verifies that a seller’s trading decision is optimal. Namely, if the net utility \(p(t) - \Delta V_\ell(t)\) of selling is strictly positive, then \(u_\ell(t) = \rho \mu_{\ell o}(t)\), meaning that all \(\ell o\) investors in contact with marketmakers choose to sell. If, on the other hand, the net utility of selling is zero, then \(\ell o\) investors are indifferent between selling and not selling. As a result, \(u_\ell(t) \leq \rho \mu_{\ell o}(t)\), meaning that some \(\ell o\) investors might choose not to sell.

### C.2 Proof of Theorem 2 and Proposition 6

#### C.2.1 Solution method

The idea is to identify equilibrium objects with the multipliers of Appendix B.4, as in Table 3. For each \(m \in [0, m^*]\), let \(\Lambda(m)\) be the (set of) multipliers associated with the buffer allocation \(m\). If \(m \in [0, m^*]\), or if \(m = m^*\) and \(m^* < \bar{m}\), then \(\Lambda(m)\) is a singleton. If \(m = m^*\) and \(m^* = \bar{m}\), then \(\Lambda(m^*)\) is a set (see Case 2 of Appendix B.4). Let \(\bar{\lambda}\) (respectively \(\lambda\)) be the element of \(\Lambda(m^*)\) with largest (smallest) \(\lambda_I(t_{m^*})\). By construction, \(\bar{\lambda}\) has no jump at \(t = t_m\). Lastly, one lets \(\bar{a}_0^* \equiv \Delta_I(t_{m^*}) e^{-rt_{m^*} m^*}\) and \(\bar{a}_0^* \equiv \bar{\lambda}_I(t_{m^*}) e^{-rt_{m^*} m^*}\). By construction, \(a_0^* = \lim_{m \to m^*} \lambda_I(t_m) e^{-rt_m m}\). Moreover, \(\bar{a}_0^* \leq \bar{a}_0^*\), with an equality if \(m^* < \bar{m}\).

Now, one can construct a competitive equilibrium using the following “backsolving” method. First, one picks some buffer allocation \(m \in [0, m^*]\) and multipliers \(\lambda \in \Lambda(m)\). Then, given \(\lambda\), one guesses that price and values are given as in Table 3. If \(m < m^*\), or if \(m = m^*\) and \(\lambda \neq \bar{\lambda}\), one takes time-zero capital to be \(a(0) = \lambda_I(t_m) e^{-rt_m m}\). If \(m = m^*\) and \(\lambda = \bar{\lambda}\), one can take any \(a(0) \in [\bar{a}_0^*, \infty)\). Subsection C.2.2 verifies that, given this time-zero capital, the buffer allocation, the price, and the values are the basis of a competitive equilibrium.

\(^{25}\)Because of linear utility, it also shows that selling one share is always weakly preferred to selling a smaller quantity \(q \in [0, 1]\).
Conversely, let’s consider any \( a(0) \in [0, a_0^*] \). Given that \( m \mapsto \lambda_I(t_m^*)e^{-rt_m^*}m \) is continuous and is zero at \( m = 0 \), the construction of the previous paragraph implies that there exists a competitive equilibrium implementing a buffer allocation with some inventory bound \( m \in [0, m^*] \). For any \( a(0) \in [a_0^*, \infty) \), the previous paragraph implies that there exists a competitive equilibrium implementing the buffer allocation with maximum inventory \( m^* \). In particular, if \( a(0) \geq a_0^* \), the multipliers do not jump at time \( t_m^* \). This establishes Proposition 6 and Theorem 2.

C.2.2 Verification

The current value Lagrangian for the representative marketmaker’s problem is

\[
\mathcal{L}(t) = c(t) + \hat{\lambda}_I(t)(u_{\ell}(t) - u_h(t)) \\
+ \lambda_a(t)(ra(t) + p(t)(u_h(t) - u_{\ell}(t)) - c(t)) \\
+ \eta_I(t)I(t) + \eta_a(t)a(t) + \hat{w}_c(t)c(t).
\]

The first-order sufficient conditions are

\[
1 + \hat{w}_c(t) = \hat{\lambda}_a(t) \tag{98}
\]
\[
\hat{\lambda}_I(t) = \hat{\lambda}_a(t)p(t) \tag{99}
\]
\[
\lambda_I(t) = \eta_I(t) + \hat{\lambda}_I(t) \tag{100}
\]
\[
\hat{\lambda}_a(t) = -\eta_a(t) \tag{101}
\]
\[
\hat{w}_c(t) \geq 0 \quad \text{and} \quad \hat{w}_c(t)c(t) = 0 \tag{102}
\]
\[
\eta_I(t) \geq 0 \quad \text{and} \quad \eta_I(t)I(t) = 0 \tag{103}
\]
\[
\eta_a(t) \geq 0 \quad \text{and} \quad \eta_a(t)a(t) = 0 \tag{104}
\]
\[
\hat{\lambda}_a(t^+) - \hat{\lambda}_a(t^-) \leq 0 \quad \text{if} \ a(t) = 0, \tag{105}
\]

together with the transversality conditions

\[
\lim_{t \to +\infty} \lambda_x(t)x(t)e^{-rt} = 0, \tag{106}
\]

for \( x \in \{I, a\} \). The Bellman equations and optimality conditions for the investors are (90)-(97). Direct comparison shows that a solution of the system (98)-(104), (106), (90)-(97) of equilibrium equations is \( p(t) = \lambda_I(t) \),

\[
\hat{\lambda}_a(t) = 1 + \left( \frac{\lambda_I(t_m^+)}{\lambda_I(t_m^-)} - 1 \right) \mathbb{1}_{\{t < t_m\}},
\]

\[
\hat{\eta}_a(t) = 0, \quad \hat{\lambda}_I(t) = \hat{\lambda}_a(t)\lambda_I(t), \quad \hat{\eta}_I(t) = \hat{\lambda}_a(t)\eta_I(t), \quad \hat{w}_c(t) = \hat{\lambda}_a(t) - 1, \quad \Delta V_\ell(t) = \lambda_\ell(t), \quad \text{and} \quad \Delta V_h(t) = \lambda_h(t),
\]

together with the corresponding inventory-constrained allocation, and some consumption process \( c(t) \) such that

\[
c(t) = 0 \quad \text{for} \quad t \leq t_2 \tag{107}
\]
\[
c(t) = ra(t_2) \quad \text{for} \quad t \geq t_2. \tag{108}
\]

To conclude the optimality verification argument for a marketmaker, one needs to check the jump condition (105), and that \( a(t) \geq 0 \) for all \( t \geq 0 \). To that end, one notes that, for \( t \in [t_1, t_2] \), \( d/dt(a(t)e^{-rt}) = -p(t)\tilde{I}(t) \) and \( \tilde{I}(t_m) = 0 \). This implies that \( a(t)e^{-rt} \) is continuously differentiable and achieves its minimum at \( t = t_m \). By construction \( a(t_m) = 0 \).
References


Greenberg, M. (2003): “Shake up the NYSE Specialist System or Drop it,” Financial Times. 4


Parry, R. T. (1997): “The October ’87 Crash Ten Years Later,” Federal Reserve Bank of San Francisco Economic Letter, 97-32. 4, 8


