Leaning against the wind

Pierre-Olivier Weill*

February 20, 2005

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, 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, 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, and the Federal Reserve Bank of Atlanta. 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 [2003]) 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

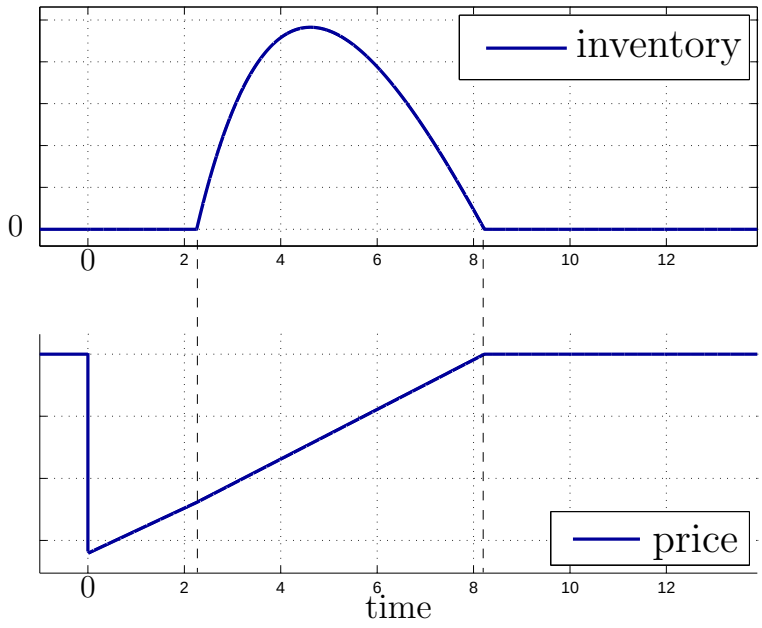

Figure 1: Features of the Competitive Equilibrium.

sufficient capital (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.[2] 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.

[2]For instance, the glossary of `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 loose 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. Instead, the social benefit of marketmakers' liquidity provision is to share risk with the sellers before the arrival of buyers. In the present model, the social benefit of liquidity provision is not to share risk but 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 [1983]. 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 [2001b], Weill [2002], Vayanos and Wang [2003], Spulber [1996] and Hall and Rust [2003]. The present model builds specifically on the work of Duffie, Gârleanu and Pedersen [2001a]. 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. The second Section describes the economic environment. The third Section solves for socially optimal dynamic liquidity provision. The fourth Section studies the implementation of this optimum in a competitive equilibrium. The fifth Section introduces borrowing-constrained marketmakers, the sixth Section discusses policy implications, and the last Section 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 $\theta(t) = 1 - \delta_2$, for some $\delta_2 \in (0, 1)$ 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 $\theta(t) = 1 - \delta_1$, for some $\delta_1 \in (0, \delta_2)$. Investors transit randomly, and pair-wise independently, from low to high marginal utility with intensity[4] γ_u, and from high to low marginal utility with intensity γ_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. The assumption that $\delta_1 < \delta_2$ means 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.

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 $I(t)$

[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 $\theta(t) = 1 - \delta_1$, the time $\inf\{u \geq 0 : \theta(t+u) \neq \theta(t)\}$ until the next switch is exponentially distributed with parameter γ_u. The successive switching times are independent.

of a representative marketmaker satisfies the shortselling constraint[5]

$$I(t) \geq 0. \tag{1}$$

An investor also cannot shortsell 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 "ℓ") and his ownership status (owner of one unit, "o," or non-owner, "n"). The set of investors' types is $\mathcal{T} \equiv \{\ell o, hn, ho, \ell n\}$. In anticipation of their equilibrium behavior, low-marginal-utility owners (ℓo) are named "sellers," and high-marginal-utility non-owners (hn) are "buyers." For each $\sigma \in \mathcal{T}$, $\mu_\sigma(t)$ denotes the fraction of type-σ investors in the total population of investors. These fractions must satisfy two accounting identities. First, of course,

$$\mu_{\ell o}(t) + \mu_{hn}(t) + \mu_{\ell n}(t) + \mu_{ho}(t) = 1. \tag{2}$$

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

$$\mu_{ho}(t) + \mu_{\ell o}(t) + I(t) = s. \tag{3}$$

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 $\mu_h(t) \equiv \mu_{ho}(t) + \mu_{hn}(t)$ of high-marginal-utility investors solves the ordinary differential equation (ODE)

$$\begin{aligned} \dot{\mu}_h(t) &= \gamma_u\big(\mu_{\ell o}(t) + \mu_{\ell n}(t)\big) - \gamma_d\big(\mu_{ho}(t) + \mu_{hn}(t)\big) \\ &= \gamma_u\big(1 - \mu_h(t)\big) - \gamma_d \mu_h(t) \\ &= \gamma_u - \gamma \mu_h(t), \end{aligned} \tag{4}$$

[5]One can allow for a limited amount of shortselling by letting $I(t) \geq -\varepsilon$ for some small $\varepsilon \in \mathbb{R}_+$. Then the results of this paper would continue to hold, with the change of variable $\tilde{I}(t) = I(t) + \varepsilon$ and $\tilde{s} = s + \varepsilon$.

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 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

$$\mu_h(t) = y\left(1 - e^{-\gamma t}\right), \tag{5}$$

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

$$s < y. \tag{6}$$

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.

$\mu_{\ell o}(0)$	$\mu_{hn}(0)$	$\mu_{\ell n}(0)$	$\mu_{ho}(0)$	$I(0)$
s	0	$1-s$	0	0

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 $\rho > 0$, where $\rho \neq \gamma$. 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

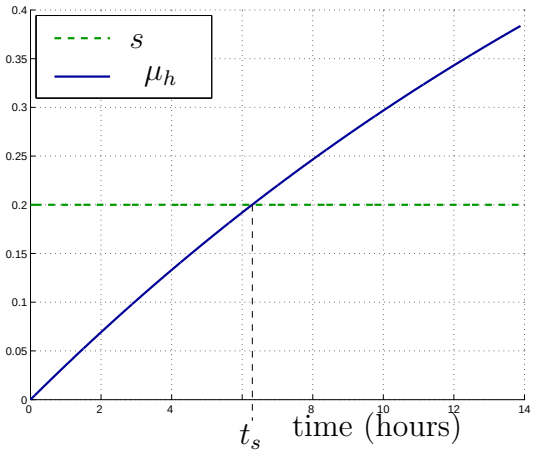

Figure 2: Dynamic of $\mu_h(t)$.

earlier) implies that contacts between type-σ 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-σ investors.

Alternative Specification of the Contact-time Technology

In the present model, an investor's average contact time $1/\rho$ with marketmakers is constant over time and does not depend on the fraction of buyers and sellers in the market. One could study an alternative model in which the average contact time would increase when the selling or the buying pressure is larger, representing for instance congestions. For example, one could assume that the instantaneous rate of contact with type-σ investors is no longer the linear function $\rho\mu_\sigma(t)$ but instead is an increasing and strictly concave function of $\mu_\sigma(t)$. This alternative non-linear specification is much less tractable and requires a numerical solution method. Moreover, the basic intuitions for the welfare improving role of marketmakers' liquidity provision would be the same as with the present linear specification. An interesting difference is that the socially-optimal allocation would need to be implemented in a "competitive-search" equilibrium, along the lines of Moen [1997], Shimer [1995], and Mortensen and Wright [2002].

Interpreting Random Contact Times

The random contact times represent a broad range of execution delays, includ-

Table 2: Parameter Values.

Parameters		Value
Measure of Shares	s	0.2
Discount Rate	r	5%
Contact Intensity	ρ	1000
Intensity of Switch to High	γ_u	90
Intensity of Switch to Low	γ_d	10
Low marginal utility	$1-\delta_1$	0.01
Marketmaker marginal utility	$1-\delta_2$	0

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.

ing the time to contact a marketmaker, to negotiate and process an order, to deliver an asset, or to transfer a payment. The parameter ρ 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] 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. 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." Along similar lines, Wigmore [1998] argues that the specialist system was the "weak link" of the October 1987 crash, because it was not designed to handle the massive selling pressure in a timely fashion.

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 deal-

ers were forced to manually execute some customers orders."[6] This suggests that technological improvements which followed the 1987 crash did not prevent substantial order-execution delays from arising during the crash of 1997.

Market Orders versus Limit Orders

In what follows, the trades of investors with marketmakers are interpreted as market orders. This means in particular that investors cannot trade with limit orders. This might be viewed as a strong assumption since limit orders are often considered good substitute for the liquidity provision of marketmakers. Empirical evidence suggest however that investors do not find limit orders very attractive in bad times and that they instead prefer to directly send market orders to floor brokers who provide more flexible execution. Goldstein and Kavajecz [2003] report that, during the market break of October 1997, there was a dramatic drain of liquidity in the limit-order book. They show that, 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. The authors write in their conclusion that "the results suggest that extreme uncertainty concerning the ability to trade continuously caused market participants to change their behavior in such a way that it effectively shut down liquidity provision via the limit order book."

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 5. 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.

[6] SEC Staff Legal Bulletin No. 8, `http://www.sec.gov/interps/legal/slbmr8.htm`

For instance, the flow rate $u_\ell(t)$ of assets that a marketmaker takes from low-marginal-utility investors is subject to the order-flow constraint

$$-\rho\mu_{\ell n}(t) \le u_\ell(t) \le \rho\mu_{\ell o}(t). \tag{7}$$

The upper (lower) bound shown in (7) is the flow of ℓo (ℓn) 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

$$-\rho\mu_{ho}(t) \le u_h(t) \le \rho\mu_{hn}(t). \tag{8}$$

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 ℓo 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

$$\dot{I}(t) = u_\ell(t) - u_h(t). \tag{9}$$

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

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

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, ℓn, and ho investors are, respectively,

$$\begin{aligned}
\dot{\mu}_{hn}(t) &= -u_h(t) - \gamma_d\mu_{hn}(t) + \gamma_u\mu_{\ell n}(t) && (11)\\
\dot{\mu}_{\ell n}(t) &= u_\ell(t) - \gamma_u\mu_{\ell n}(t) + \gamma_d\mu_{hn}(t) && (12)\\
\dot{\mu}_{ho}(t) &= u_h(t) - \gamma_d\mu_{ho}(t) + \gamma_u\mu_{\ell o}(t). && (13)
\end{aligned}$$

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

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

(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 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 Pareto optimal allocation must maximize

$$\int_0^{+\infty} e^{-rt}\Big(\mu_{ho}(t) + (1-\delta_1)\mu_{\ell o}(t) + (1-\delta_2)I(t)\Big)dt, \tag{14}$$

the equally weighted sum of investors' and marketmakers' intertemporal utilities. This criterion is deterministic, reflecting pairwise independence of investors' marginal-utility and contact-time processes. Conversely, an asset allocation maximizing (14) is Pareto optimal.[7] This discussion motivates the following definition of an optimal allocation.

Definition 2 (Optimal Allocation.) *An 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

$$\mu_{\ell o}(t) = s - \mu_h(t) + \mu_{hn}(t). \tag{15}$$

[7] Weill [2004] shows that this result also hold under the alternative assumption that cash transfers are be made dynamically over time, subject to the same trading technology as asset transfers.

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

$$\rho\mu_{hn}(t) < \rho\mu_{\ell o}(t) \tag{16}$$

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 2ε can be made arbitrarily small.

The benefit is that, for a sufficiently small ε, 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 γ_u, or (ii) the seller establishes another contact with a marketmaker with intensity ρ. 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 δ_1, 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 $\delta_2 - \delta_1$. If ε 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*[8]

$$u_\ell(t) = \rho\mu_{hn}(t)\mathbb{I}_{[0,t_1)}(t) + \rho\mu_{\ell o}(t)\mathbb{I}_{[t_1,+\infty)}(t) \tag{17}$$

$$u_h(t) = \rho\mu_{hn}(t)\mathbb{I}_{[0,t_2)}(t) + \rho\mu_{\ell o}(t)\mathbb{I}_{[t_2,+\infty)}(t) \tag{18}$$

$$I(t_2) = 0. \tag{19}$$

[8]In what follows, $\mathbb{I}_A(\,\cdot\,)$ denotes the indicator function of some set $A \subseteq \mathbb{R}$.

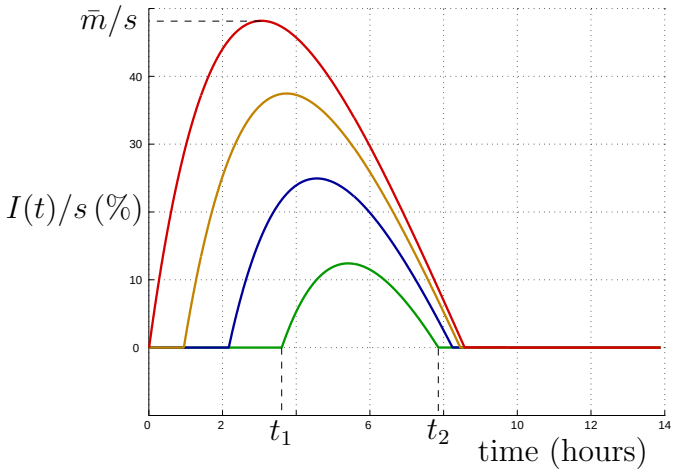

Figure 3: Illustrative buffer allocations.

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 $u_\ell(t) = \rho\mu_{\ell o}(t)$ or $u_h(t) = \rho\mu_{hn}(t)$. Because of the linear objective (14), it is natural to guess that an optimal allocation will also have this bang-bang property. In the next subsection, Theorem 1 will confirm this conjecture, showing that the optimal allocation belongs to the family of buffer allocations.

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 exists some $\bar{m} \in \mathbb{R}_+$, some strictly decreasing function $\psi : [0, \bar{m}] \to \mathbb{R}_+$, and some strictly increasing 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),

$$\begin{aligned} m &= \max_{t \in \mathbb{R}_+} I(t) && (20) \\ \psi(m) &= \arg\max_{t \in \mathbb{R}_+} I(t) && (21) \\ t_1 &= \psi(m) - \phi_1(m) && (22) \\ t_2 &= \psi(m) + \phi_2(m), && (23) \end{aligned}$$

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.

$$\lim_{m \to 0^+} \frac{W(m) - W(0)}{m} > 0. \tag{24}$$

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 $\big(\mu^m(t), I^m(t), u^m(t)\big)$ with maximum inventory position m, I define the cost of holding inventory as

$$C(m) \equiv \int_0^{+\infty} e^{-rt}(\delta_2 - \delta_1) I^m(t)\, dt, \tag{25}$$

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

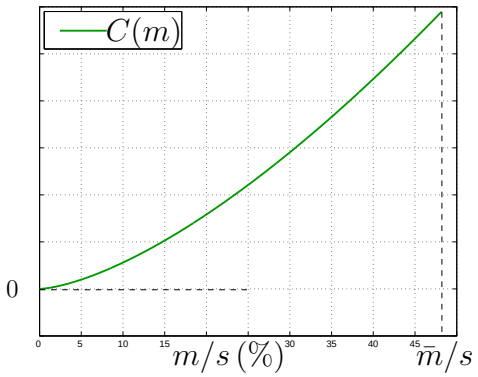

Figure 4: The intertemporal cost of accumulating inventory.

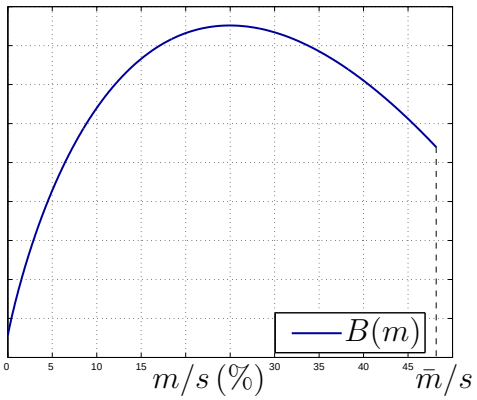

Figure 5: The intertemporal benefit of accumulating inventory.

inventory is increasing in the maximum inventory position. Intuitively, an additional unit of inventory is transferred later in time, implying that the holding cost $\delta_2 - \delta_1$ is incurred for a longer time period. Similarly, I define the benefit of holding inventory as

$$B(m) \equiv W(m) + C(m). \tag{26}$$

The function $B(m)$ is a measure of social welfare which is compensated for the holding cost $\delta_2 - \delta_1$ 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 ℓo 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 expansive 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 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 optimal allocation.

First-Order Sufficient Conditions

The first-order sufficient conditions are based on Seierstad and Sydsæter [1977], and are described in detail in Appendix E. The accounting identities $\mu_{ho}(t) = \mu_h(t) - \mu_{hn}(t)$ 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 $\big(\mu_{\ell o}(t), \mu_{hn}(t), I(t)\big)$. The "current-value" Lagrangian (see Kamien and Schwartz [1991], Part II, Section 8) is

$$\begin{aligned}\mathcal{L}(t) = \quad & \mu_h(t) - \mu_{hn}(t) + (1-\delta_1)\mu_{\ell o}(t) + (1-\delta_2)I(t) \\ & + \lambda_{\ell o}(t)\left(-u_\ell(t) - \gamma_u\mu_{\ell o}(t) - \gamma_d\mu_{hn}(t) + \gamma_d\mu_h(t)\right) \\ & + \lambda_{hn}(t)\left(-u_h(t) - \gamma_u\mu_{\ell o}(t) - \gamma_d\mu_{hn}(t) + \gamma_u(1-\mu_h(t))\right) \\ & + \lambda_I(t)\left(u_\ell(t) - u_h(t)\right) \\ & + w_{\ell o}(t)\left(\rho\mu_{\ell o}(t) - u_\ell(t)\right) + w_{hn}(t)\left(\rho\mu_{hn}(t) - u_h(t)\right) + \eta_I(t)\, I(t).\end{aligned} \tag{27}$$

The multiplier $\lambda_{\ell o}(t)$ of the ODE (10) represents the social value of increasing the flow of investors from the ℓn type to the ℓo type or, equivalently, the value of transferring an asset to an ℓn investor. One gives a similar interpretation to the multipliers $\lambda_{hn}(t)$ and $\lambda_I(t)$ of the ODEs (11) and (9), respectively. The multipliers $w_{\ell o}(t)$ and $w_{hn}(t)$ of the flow constraints (7) and (8) represent the social value of increasing the rate of contact with ℓo and hn investors,

respectively.[9] The multiplier on the shortselling constraint (1) is $\eta_I(t)$. The first-order condition with respect to the controls $u_\ell(t)$ and $u_h(t)$ are

$$\begin{aligned} w_{\ell o}(t) &= -\lambda_{\ell o}(t) + \lambda_I(t) && (28)\\ w_{hn}(t) &= -\lambda_{hn}(t) - \lambda_I(t), && (29) \end{aligned}$$

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

$$w_{\ell o}(t) \geq 0 \quad \text{and} \quad w_{\ell o}(t)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) = 0, \tag{30}$$

and

$$w_{hn}(t) \geq 0 \quad \text{and} \quad w_{hn}(t)\big(\rho\mu_{hn}(t) - u_h(t)\big) = 0. \tag{31}$$

The multipliers $w_{\ell o}(t)$ and $w_{hn}(t)$ are non-negative because a marketmaker can ignore additional contacts. The complementary-slackness condition (30) means that, when the marginal value $w_{\ell o}(t)$ of additional contact is strictly positive, a marketmaker should take the assets of all ℓo 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 $\lambda_{\ell o}(t)$, $\lambda_{hn}(t)$, and $\lambda_I(t)$ are

$$\begin{aligned} r\lambda_{\ell o}(t) &= 1 - \delta_1 + \gamma_u(-\lambda_{hn}(t) - \lambda_{\ell o}(t)) + \rho w_{\ell o}(t) + \dot{\lambda}_{\ell o}(t) && (33)\\ r\lambda_{hn}(t) &= -1 - \gamma_d(\lambda_{hn}(t) + \lambda_{\ell o}(t)) + \rho w_{hn}(t) + \dot{\lambda}_{hn}(t) && (34)\\ r\lambda_I(t) &= 1 - \delta_2 + \eta_I(t) - \eta_M(t) + \dot{\lambda}_I(t), && (35) \end{aligned}$$

respectively. For instance, (33) decomposes the flow value $r\lambda_{\ell o}(t)$ of transferring an asset to a low-marginal-utility investor. The first term, $1 - \delta_1$, is

[9] 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 flow value of the dividend, for a low-marginal-utility investor. The second term, $\gamma_u\big(-\lambda_{hn}(t)-\lambda_{\ell o}(t)\big)$, is the expected rate of net utility associated with a transition to high marginal utility. That is, with intensity γ_u, $\lambda_{\ell o}(t)$ becomes the value $-\lambda_{hn}(t)$ of transferring an asset to a high-marginal-utility investor. The third term, $\rho w_{\ell o}(t)$, is the expected rate of net utility of a contact between an ℓo investor and a marketmaker. The multipliers $\big(\lambda_{\ell o}(t), \lambda_{hn}(t), \lambda_I(t)\big)$ must satisfy the following additional restrictions. First, they must satisfy the transversality conditions[10]

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

for $j \in \{\ell o, hn, I\}$. Second, the multipliers $\lambda_{hn}(t)$ and $\lambda_{\ell o}(t)$ 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 ε, improving the objective by $\varepsilon\big(\lambda_I(t^+) - \lambda_I(t^-)\big)e^{-rt}$.

The Optimal Allocation

Appendix B guesses and verifies that the (essentially unique) 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 $w_{\ell o}(t_1) = 0$ is used to find the breaking-times t_1 and t_2.

Theorem 1 (Optimal Allocation.) *There exists an optimal allocation* $\big(\mu^*(t), I^*(t), u^*(t), t \geq 0\big)$. *This allocation is a buffer allocation with breaking times* (t_1^*, t_2^*) *determined by*

$$e^{-\gamma t_1^*} = \left(1 - \frac{s}{y}\right) \frac{1 - e^{-\rho\Delta^*}}{\rho} \frac{\rho - \gamma}{e^{-\gamma\Delta^*} - e^{-\rho\Delta^*}} \tag{38}$$

$$t_2^* = t_1^* + \Delta^* \tag{39}$$

$$\Delta^* = \min\left\{\frac{1}{r+\rho}\log\left(1 + \frac{\delta_1(r+\rho)}{\delta_2\gamma_u + (\delta_2 - \delta_1)(r+\rho+\gamma_d)}\right), \bar{\Delta}\right\},$$

[10]This condition, derived in Appendix E, allows to complete the standard optimality verification argument when the time horizon is infinite.

where $\bar{\Delta} \equiv \phi_1(\bar{m}) + \phi_2(\bar{m})$.

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 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 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* $\big(\mu(t), I(t), u(t)\big)$ *is an 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 Δ^* of the inventory-accumulation period. Another measure would be the maximum inventory position m^*, which is strictly monotonic with Δ^*, provided that ρ and γ are held constant.[11] With Theorem 1, Δ^* can be written as $F(\rho, r, \delta_1, \delta_2, \gamma_u, \gamma_d)$, for some continuous function $F(\,\cdot\,)$. The following proposition provides some natural comparative statics.

Proposition 4 (A Comparative Static.) *Let* $x = (\rho, r, \delta_1, \delta_2, \gamma_u, \gamma_d)$ *be a vector of exogenous parameters. If* $F(x) < \bar{\Delta}$*, then* $F(\,\cdot\,)$ *is differentiable at* x*, with partial derivatives*

$$\frac{\partial F}{\partial \rho} < 0, \ \frac{\partial F}{\partial r} < 0, \ \frac{\partial F}{\partial \delta_1} > 0, \ \frac{\partial F}{\partial \delta_2} < 0, \ \frac{\partial F}{\partial \gamma_d} < 0, \text{ and } \frac{\partial F}{\partial \gamma_u} < 0. \tag{40}$$

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

[11] 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 ρ and γ.

depend on (r, δ_1, δ_2). Proposition 4 shows that the inventory-accumulation period is longer when an investor's holding cost δ_1 is larger. It is shorter whenever the order-execution technology is faster, the agents are more impatient, marketmakers holding cost is larger, or investors' switching intensities γ_d and γ_u are larger. This last comparative static reflects the fact that larger γ_d and γ_u reduce the net utility of transferring the asset. Namely, with a larger γ_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 γ_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 optimal allocation as ρ 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 Δ^* 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* $(r, \delta_1, \delta_2, \gamma_u, \gamma_d)$, *as* $\hat{\rho} \to +\infty$,

$$F(\hat{\rho}, r, \delta_1, \delta_2, \gamma_u, \gamma_d) \to 0 \tag{41}$$

$$(\phi_1 + \phi_2)^{-1} \circ F(\hat{\rho}, r, \delta_1, \delta_2, \gamma_u, \gamma_d) \to 0. \tag{42}$$

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 optimal allocation can be implemented in a competitive equilibrium.

4.1 Competitive Marketmakers

This subsection describes a competitive market structure which implements the optimal allocation. Weill [2004] shows that the efficiency result generalizes to environments with stochastic contact intensity ρ and transition intensities γ_u and γ_d.

Marketmaker's Problem

A marketmaker has access to a 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 $\big\{p(t), t \geq 0\big\}$. Hence, her wealth $a(t)$ and her inventory position $I(t)$ evolve according to

$$\begin{aligned} \dot{a}(t) &= ra(t) + (1-\delta_2)I(t) + p(t)\big(u_h(t) - u_\ell(t)\big) - c(t) && (43) \\ \dot{I}(t) &= u_\ell(t) - u_h(t). && (44) \end{aligned}$$

In addition, she faces the borrowing and shortselling constraints

$$\begin{aligned} a(t) &\geq 0 && (45) \\ I(t) &\geq 0. && (46) \end{aligned}$$

Lastly, it is assumed that at time zero, a marketmaker's holds no inventory ($I(0) = 0$) and maintains a strictly positive amount of capital $a(0)$. This Section 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.) The marketmaker's objective is to maximize the present value of her consumption stream,

$$\int_0^{+\infty} e^{-rt} c(t)\, dt, \qquad (47)$$

with respect to $\big\{a(t), I(t), u_\ell(t), u_h(t), c(t), t \geq 0\big\}$, subject to the constraints (43)-(46), and the constraint that $u_\ell(t)$ and $u_h(t)$ are piecewise continuous. The above formulation will imply 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.

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 the price $p(t)$. I solve the investor's problem using a "guess and verify" method.[12] Specifically, I guess that, in equilibrium, an ℓo (hn) investor always finds it weakly optimal to sell (to buy). If an ℓo (hn) investor is indifferent between selling and not selling, he might choose not to sell (to buy). Lastly, I guess that investors of types ℓn and ho never trade. The time-t continuation utility of an investor of type $\sigma \in \mathcal{T}$ who follows this policy is denoted $V_\sigma(t)$. Hence, a seller's reservation value is

$$\Delta V_\ell(t) \equiv V_{\ell o}(t) - V_{\ell n}(t). \tag{48}$$

Similarly, a buyer's reservation value is $\Delta V_h(t) \equiv V_{ho}(t) - V_{hn}(t)$. The reservation value $\Delta V_\ell(t)$ of a seller solves

$$r\Delta V_\ell(t) = 1 - \delta_1 + \gamma_u\big(\Delta V_h(t) - \Delta V_\ell(t)\big) + \rho\big(p(t) - \Delta V_\ell(t)\big) + \Delta\dot{V}_\ell(t), \tag{49}$$

Equation (49) breaks the reservation value into four terms. The first term, $1 - \delta_1$, is net rate of dividends for a low-marginal-utility investor. The second term, $\gamma_u\big(\Delta V_h(t) - \Delta V_\ell(t)\big)$, is the expected rate of net utility associated with transition to the high-marginal-utility state, because, with intensity γ_u, the investor's reservation value changes from $\Delta V_\ell(t)$ to $\Delta V_h(t)$. The third term, $\rho\big(p(t) - \Delta V_\ell(t)\big)$, is the expected rate of net utility associated with selling the asset, and the fourth term reflects time variation in the reservation value. Similarly the reservation value $\Delta V_h(t)$ of a buyer solves

$$r\Delta V_h(t) = 1 + \gamma_d\big(\Delta V_\ell(t) - \Delta V_h(t)\big) - \rho\big(\Delta V_h(t) - p(t)\big) + \Delta\dot{V}_h(t), \tag{50}$$

[12] Appendix D describes the investor's dynamic programming problem in details. In particular, it is shown that, when he is permitted to hold any quantity $q \in [0, 1]$ of shares, an investor finds it optimal to hold either the minimum quantity of zero unit, or the maximum quantity of one unit.

Lastly, I impose the transversality conditions

$$\lim_{t\to+\infty} \Delta V_j(t) e^{-rt} = 0, \tag{51}$$

for $j \in \{h, \ell\}$. This transversality condition allows to complete the optimality verification argument described in detail in Appendix D.

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

(i) Given the price $p(t)$*,* $\big(I(t), u(t), c(t), a(t)\big)$ *solves the marketmaker's problem.*

(ii) Given the price $p(t)$*, the reservation values* $\big(\Delta V_\ell(t), \Delta V_h(t)\big)$ *solve equations (49)-(51) and satisfy, at each time,*

$$p(t) - \Delta V_\ell(t) \geq 0 \tag{52}$$
$$\Delta V_h(t) - p(t) \geq 0 \tag{53}$$
$$\big(p(t) - \Delta V_\ell(t)\big)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) = 0 \tag{54}$$
$$\big(\Delta V_h(t) - p(t)\big)\big(\rho\mu_{hn}(t) - u_h(t)\big) = 0. \tag{55}$$

Equations (52) through (55) verify the optimality of investors' policies. For instance, equation (52) means that the net utility of selling is positive, which verifies that a seller ℓo finds it weakly optimal to sell. Equation (54), 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 ℓo investors in contact with marketmakers choose to sell. If, on the other hand, the net utility of selling is zero, then ℓo investors are indifferent between selling and not selling. As a result, $u_\ell(t) \leq \rho\mu_{\ell o}(t)$, meaning that some ℓo investors might choose not to sell.

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

Theorem 2 states that the optimal allocation can be implemented in some competitive equilibrium. The proof identifies the price and the reservation

Table 3: Identifying Prices with Multipliers.

Equilibrium Objects	Multipliers	Constraints
$p(t)$	$\lambda_I(t)$	$\dot{I}(t) = u_\ell(t) - u_h(t)$
$\Delta V_\ell(t)$	$\lambda_{\ell o}(t)$	$\dot{\mu}_{\ell o}(t) = -u_\ell(t) \dots$
$p(t) - \Delta V_\ell(t)$	$w_{\ell o}(t)$	$u_\ell(t) \leq \rho \mu_{\ell o}(t)$
$\Delta V_h(t)$	$-\lambda_{hn}(t)$	$\dot{\mu}_{hn}(t) = -u_h(t) \dots$
$\Delta V_h(t) - p(t)$	$w_{hn}(t)$	$u_h(t) \leq \rho \mu_{hn}(t)$

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).

values with the Lagrange multipliers of the optimal allocation (see Table 3). For instance, the asset price $p(t)$ is equal to the multiplier $\lambda_I(t)$ 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.

The minimum capital $\underline{a}_0$ can be shown to be $\max_{t \in \mathbb{R}_+} \lambda_I(t) I(t) e^{-rt}$. In other words, optimal implementation is obtained if there is enough market-making capital to purchase the inventory position of maximum present value.

Equilibrium Price Path and Marketmaker's 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 and increases thereafter.[13] The price has three phases reflecting the three phases of the socially-optimal allocation. Before the first breaking time t_1^*, marketmakers do not accommodate the selling pressure. This implies that the price must adjust in order to make a seller indifferent between selling or not, meaning that $p(t) = \Delta V_\ell(t)$. In between the two breaking times t_1^* and t_2^*, marketmakers accommodate all

[13] 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 κ. One can show that the results of this paper would apply, provided that either κ is small enough, or $t_1^* > 0$. For Figure 6, it is assumed that $\kappa = 0$.

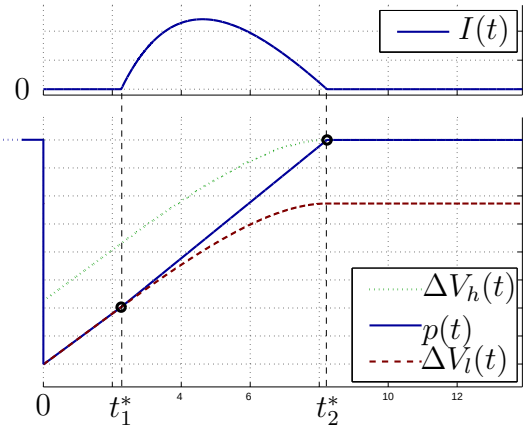

Figure 6: The Equilibrium Price Path.

of the selling pressure. As a result, the "marginal investor" is a marketmaker, and

$$\Delta V_\ell(t) < p(t) < \Delta V_h(t). \tag{56}$$

In particular, $p(t) > \Delta V_\ell(t)$, meaning that the liquidity provision of marketmakers raises the asset price above a seller's marginal utility. In that sense, marketmakers' liquidity provision helps to "support" the price level. After the second breaking time t_2^*, the asset price has reached its steady-state value.

A marketmaker buys the asset early, at a low price, and later re-sells at a higher price. The price growth rate in $[t_1^*, t_2^*]$ exactly compensates a marketmaker for the time value of cash, net of the flow value $1-\delta_2$ of the dividend. In other words, the 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 price growth rate is lower, making it unprofitable for the marketmaker to buy the asset on her own account.

The previous paragraph implies in particular that a marketmaker intertemporal utility is equal to $a(0)$, the present 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.

Allocative Efficiency in Related Search-and-Matching Models

Hosios [1990] provides a necessary condition for allocative efficiency in search-and-matching models: the buyer's (seller's) bargaining strength should be

equal to the elasticity of the matching function[14] with respect to the mass of buyers (sellers). This condition is satisfied in the present model, provided that one recognizes that there are two matching functions: the rate $\rho\mu_{hn}(t)$ of contact between buyers and marketmakers, and the rate $\rho\mu_{\ell o}(t)$ of contact between sellers and marketmakers. Because these two matching functions have unit elasticity, the Hosios conditions would prescribe that the bargaining strengths of buyers and sellers should be both equal to 1. In particular, in order to achieve efficiency, marketmakers should buy and sell at the same price and should make zero intertemporal profit. These two conditions are satisfied by the equilibrium of Theorem 2.

5 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 Section describes an amplification mechanism associated with insufficient capital and binding borrowing constraints. Specifically, it 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.

Because time-zero capital represents a marketmaker's purchasing power, it resembles an upper bound on a marketmaker's inventory. This Section formalizes this intuition. First, it defines, and solves for, constrained-optimal allocations, subject to an exogenous inventory upper bound. Then, it implements each of these allocations with an appropriately chosen time-zero capital.

5.1 Constrained-Optimal Allocations

It is assumed that each marketmaker faces the inventory bound

$$I(t) \leq M, \tag{57}$$

[14]The "matching function" specifies the rate of contact between buyers and sellers as a function of two arguments, the mass of buyers and the mass of sellers.

for some $M \in \mathbb{R}_+$. Subject to this additional constraint, one can reproduce the analysis of the previous Section.

Definition 5 (Constrained-Optimal Allocations.) *A constrained-optimal allocation with inventory bound $M \in \mathbb{R}_+$ is a feasible allocation maximizing (14), subject to the inventory constraint (57).*

As before, m^* denotes the maximum inventory position in the optimal allocation of Theorem 1.

Proposition 6. *There exists a constrained-optimal allocation with inventory bound M. It is a buffer allocation with maximum inventory position $m = \min\{M, m^*\}$.*

When $M \leq m^*$, the inventory constraint binds, meaning that marketmakers should accumulate as much inventory as permitted by (57). Moreover, in that case, the constraint binds only once, meaning that, given an inventory bound of $M \leq m^*$, marketmakers should maximally delay inventory accumulation. These features are illustrated in Figure 7. Importantly, Appendix B.2 shows that a binding inventory constraint produces a jump in the path of the multiplier $\lambda_I(t)$ associated with the ODE $\dot{I}(t) = u_\ell(t) - u_h(t)$. Namely, if $M < m^*$, then

$$\lambda_I(t_m^+) - \lambda_I(t_m^-) > 0, \tag{58}$$

when $I(t_m) = M$. This jump reflects the benefit, when the inventory constraint binds, of accumulating additional inventory.

5.2 Market Equilibrium

This subsection implements each constrained-optimal allocation with an appropriately chosen time-zero capital. As in the previous Section, the equilibrium price and values serve as Lagrange multipliers of the constrained-optimal allocation (see Table 3).

Proposition 7 (Implementation.) *Consider a buffer allocation with maximum inventory position $m \leq m^*$. This allocation is constrained-optimal with inventory bound $M = m$. Let t_m be the only time at which $I(t_m) = m$, and let $\lambda_I(t)$ be the Lagrange multiplier of the ODE $\dot{I}(t) = u_\ell(t) - u_h(t)$. Lastly, let $a(0) \equiv e^{-rt_m}\lambda_I(t_m^-)m$ be a marketmaker's time-zero capital. Then, there exists a competitive equilibrium whose allocation is the buffer allocation with maximum inventory position m.*

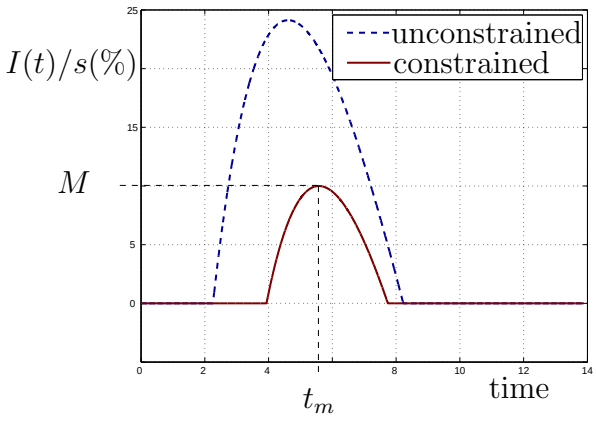

Figure 7: Constrained-optimal Allocations.

A marketmaker uses her capital to build up her inventory position between times t_1 and t_m. Time-zero capital is chosen so that $a(t_m) = 0$, meaning that a marketmaker's borrowing constraint is binding at time t_m.

As before, the equilibrium price is the multiplier $\lambda_I(t)$ of the ODE $\dot{I}(t) = u_\ell(t) - u_h(t)$. In particular, this price path features an upward jump at time t_m (see Figure 8). The price jump 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. Because of the price jump, a marketmaker makes positive profit, in that her intertemporal utility is

$$\frac{p(t_m^+)}{p(t_m^-)} a(0) > a(0) \tag{59}$$

if $p(t_m^+)/p(t_m^-) > 1$. The intuition for (59) is as follows. A marketmaker can invest her capital $a(0)$ at the risk-free rate r between time zero and time t_m. At the last instant before t_m, she spends all her capital $e^{rt_m}\, a$ in order to buy assets at price $p(t_m^-)$. She can re-sell these assets the next instant after t_m at price $p(t_m^+)$. With this trading strategy, the present value of her profit is (59). Many other trading strategies would achieve the same profit. Namely, given the equilibrium price path, the optimal timing of purchases and sales is indeterminate, as long as all assets are purchased in $[t_1, t_m]$, sold in $[t_m, t_2]$, and $a(t_m) = 0$.

Lastly, the results of Proposition 7 would also hold under the alternative assumption that marketmakers have different time-zero capital endowment, as

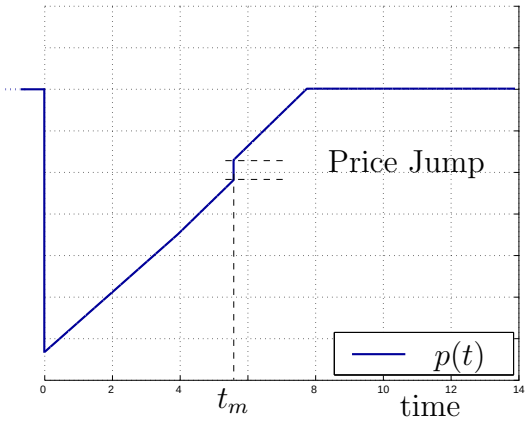

Figure 8: Equilibrium Price Path with a Small Marketmaking Capital.

long as the aggregate endowment per capita is $a(0) = \lambda_I(t_m^-)Me^{-rt_m}$.

6 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 $a\, e^{rT}$ at some time $T \geq t_2^*$. Equation (59) implies that the net present value of her profit is

$$\left(\frac{p(t_m^+)}{p(t_m^-)} - 1\right) a, \tag{60}$$

where the jump-size $\left(p(t_m^+)/p(t_m^-) - 1\right)$ depends implicitly on the time-zero aggregate marketmaking capital. 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 (60) 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 $\left(p(t_m^+)/p(t_m^-) - 1\right)$ 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. Appendix F provides an explicit model of marketmakers' borrowing limits based on some asymmetric information problems on the capital market.[15] The model supports the recommendation of subsidizing loans to marketmakers.

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.[16] 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

[15] 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.

[16] The article "Shake up the NYSE Specialist System or Drop it" was published by *The Financial Times* on October 10th, 2003.

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.[17] The present paper studies liquidity provision in terms of the Pareto criterion rather than in terms of some price smoothing objective. The 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.

[17] 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 intraday price movements."

7 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 Buffer Allocations

This Appendix studies buffer allocations and proves propositions 1 and 2.

A.1 Proof of Proposition 1

This subsection studies some features of buffer allocations. First, in any buffer allocation, the inventory position is hump-shaped. Second, it proves that the breaking times t_1 and t_2 of any buffer allocation can be viewed as functions of the maximum inventory position m. In all what follows, some buffer allocation (t_1, t_2) is fixed. For $t \in [t_1, t_2)$, the inventory position $I(t)$ evolves according to

$$\dot{I}(t) = u_\ell(t) - u_h(t) = \rho\big(\mu_{\ell o}(t) - \mu_{hn}(t)\big). \tag{61}$$

With equation (3), this ODE can be written

$$\dot{I}(t) = -\rho I(t) + \rho\big(s - \mu_h(t)\big). \tag{62}$$

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

$$\begin{aligned} H(t_1,t) &= \rho \int_{t_1}^{t} \big(s - \mu_h(z)\big) e^{\rho(z-t)} dz & (63)\\ &= (s-y)\big(1 - e^{-\rho(t-t_1)}\big) + \rho y e^{-\gamma t_1} \frac{e^{-\gamma(t-t_1)} - e^{-\rho(t-t_1)}}{\rho - \gamma}. & (64) \end{aligned}$$

The following Lemma 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.

Lemma 1 (Hump-Shaped Inventory.) *There exists a unique pair* $(t_m, t_2) \in [t_1, t_s] \times [t_s, +\infty)$ *such that*

$$\begin{aligned} \frac{\partial H}{\partial t} = 0 &\Leftrightarrow t = t_m & (65)\\ \frac{\partial H}{\partial t} > 0 &\Leftrightarrow t \in [t_1, t_m) & (66)\\ H(t_1, t_2) &= 0 & (67) \end{aligned}$$

Proof. Differentiating equation (64) shows that there is at most one $z \geq 0$ such that $\partial H/\partial t(t_1, z) = 0$. Differentiating (63) shows that $\partial H/\partial t(t_1, t_1) = \rho(s - \mu_h(t_1)) \geq 0$ and $\partial H/\partial t(t_1, t_s) = -\rho^2 \int_{t_1}^{t_s} (s - \mu_h(z)) e^{\rho(z-t)} dz \leq 0$. This implies (65) and (66). Furthermore, because $\mu_h(t) > s$ for t large enough, it follows from (63) that $H(t_1, t)$ is negative for t large enough. This, together with (65) and (66), implies that, given some $t_1 \in [0, t_s]$, there exists a unique $t_2 \in [t_s, +\infty)$ solving (67).

Equipped with this last result, one can characterize the various objects of Proposition 1. First, the maximum inventory position is defined as $m \equiv I(t_m)$. Second, one defines $t_m \equiv \psi(m)$, for some function $\psi(\,\cdot\,)$. This function can be written in closed form by substituting $\dot{I}(t_m) = 0$ and $I(t_m) = m$ in (62).

$$\psi(m) = -\frac{1}{\gamma} \log\left(1 - \frac{s-m}{y}\right). \tag{68}$$

Lastly, the breaking times (t_1, t_2) can be written $t_1 = \phi_1(m)$ and $t_2 = \phi_2(m)$, for some functions $\phi_1(\,\cdot\,)$ and $\phi_2(\,\cdot\,)$ which are characterized as follows. Solving (62) with the initial condition $I(t_m) = m$, one finds

$$I(t) = me^{-\rho(t-t_m)} + (s-y)\big(1 - e^{-\rho(t-t_m)}\big) + \rho y e^{-\gamma t_m} \frac{e^{-\gamma(t-t_m)} - e^{-\rho(t-t_m)}}{\rho - \gamma}. \tag{69}$$

Replacing (68) into (69), 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

$$G(m,z) = \left(1 + \frac{m}{y-s}\right) \frac{\rho e^{-\gamma z} - \gamma e^{-\rho z}}{\rho - \gamma}. \tag{70}$$

The following Lemma shows the existence of the functions $\phi_1(\,\cdot\,)$ and $\phi_2(\,\cdot\,)$ of Proposition 1. These can be written $\phi_i = \sqrt{\Phi_i}$, for some continuously differentiable function $\Phi_i(\,\cdot\,)$, $i \in \{1, 2\}$.

Lemma 2 (Breaking Times.) *There exists some continuously differentiable and increasing functions $\Phi_i : \mathbb{R}_+ \to \mathbb{R}_+$, $i \in \{1, 2\}$, such that $G(m, z) = 1$ if and only if $z \in \{-\sqrt{\Phi_1(m)}, \sqrt{\Phi_2(m)}\}$. Moreover $\Phi_i(0) = 0$ and $\Phi_i'(0) = 2/(\rho\gamma(y-s))$, $i \in \{1, 2\}$.*

Proof. Define, for $x \in [0, +\infty)$, the two functions $g_i(m, x) = G\big(m, (-1)^i\sqrt{x}\big)$, $i \in \{1, 2\}$. For $x > 0$, the partial derivatives of g_i with respect to x is

$$\frac{\partial g_i}{\partial x} = \left(1 + \frac{m}{y-s}\right) \frac{(-1)^i \rho\gamma}{2\sqrt{x}} \frac{e^{(-1)^{i+1}\rho\sqrt{x}} - e^{(-1)^{i+1}\gamma\sqrt{x}}}{\rho - \gamma}. \tag{71}$$

This function is strictly negative for $x > 0$. It can be extended by continuity at $x = 0$ with

$$\frac{\partial g_i}{\partial x}(m, 0) = -\left(1 + \frac{m}{y-s}\right) \frac{\rho\gamma}{2}. \tag{72}$$

Hence, $g_i(m, \,\cdot\,)$ 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 $z \in \{-\sqrt{\Phi_1(m)}, \sqrt{\Phi_2(m)}\}$.

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

$$\psi(m) - \phi_1(m) \geq 0. \tag{73}$$

The left hand side of (73) 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 $\bar{m}$.

A.2 Proof of Proposition 2

The first step is to construct multipliers for the no-inventory allocation, associated with the various constraints. The multipliers associated with the ODE of $\mu_{\ell o}(t)$, $\mu_{hn}(t)$, and $I(t)$ are denoted $\lambda_{\ell o}(t)$, $\lambda_{hn}(t)$, and $\lambda_I(t)$, respectively. The multiplier on the shortselling constraint is denoted by $\eta_I(t)$. The multipliers on the right constraints in (7) and (8) are denoted $w_{\ell o}(t)$ and $w_{hn}(t)$, respectively. These multipliers solve the ODEs

$$\begin{aligned}
\dot{\lambda}_{hn}(t) &= r\lambda_{hn}(t) + 1 + \gamma_d\big(\lambda_{hn}(t) + \lambda_{\ell o}(t)\big) - \rho w_{hn}(t) && (74)\\
\dot{\lambda}_{\ell o}(t) &= r\lambda_{\ell o}(t) - (1-\delta_1) + \gamma_u\big(\lambda_{hn}(t) + \lambda_{\ell o}(t)\big) - \rho w_{\ell o}(t) && (75)\\
\dot{\lambda}_I(t) &= r\lambda_I(t) - (1-\delta_2) - \eta_I(t) && (76)\\
w_{hn}(t) &= -\lambda_{hn}(t) - \lambda_I(t) && (77)\\
w_{\ell o}(t) &= -\lambda_{\ell o}(t) + \lambda_I(t) && (78)
\end{aligned}$$

with the complementary slackness conditions

$$w_{hn}(t)\big(\rho\mu_{hn}(t) - u_h(t)\big) = 0 \text{ and } w_{\ell o}(t)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) = 0, \tag{79}$$

and the transversality conditions $\lambda_j(t)e^{-rt} \to 0$, as $t \to \infty$, for $j \in \{hn, \ell o, I\}$. Summing (77) and (78) shows that $w_{hn}(t) + w_{\ell o}(t) = -\lambda_{hn}(t) - \lambda_{\ell o}(t)$. Summing (74) and (75) and using transversality shows that, at each time, $\lambda_{hn}(t) + \lambda_{\ell o}(t) = -\delta_1/(r+\rho+\gamma)$. Complementary slackness implies that, for $t < t_s$, $w_{\ell o}(t) = 0$ and, for $t \geq t_s$, $w_{hn}(t) = 0$. For $t \geq t_s$, transversality implies that $r\lambda_{\ell o}(t) = 1 - \delta_1(r+\gamma_d)/(r+\rho+\gamma)$ and $r\lambda_{hn}(t) = -1 + \delta_1\gamma_d/(r+\rho+\gamma)$. For $t < t_s$, one can use (74) and (75) to solve for $\lambda_{hn}(t)$ and $\lambda_{\ell o}(t)$, imposing continuity of these

multipliers at $t = t_s$. Because $\lambda_I(t) = \lambda_{\ell o}(t)$ for $t < t_s$, (75) and (76) imply that $\eta_I(t) = \delta_2 - \delta_1(r+\rho+\gamma_d)/(r+\rho+\gamma)$. Similarly, because $\lambda_I(t) = \lambda_{hn}(t)$ for $t \geq t_s$, (74) and (76) imply that $\eta_I(t) = \delta_2 - \delta_1\gamma_d/(r+\rho+\gamma)$.

Since $\lambda_I(t) = \lambda_{\ell o}(t)$ for $t < t_s$, and $\lambda_I(t) = \lambda_{hn}(t)$ for $t \geq t_s$, the multiplier $\lambda_I(\,\cdot\,)$ jumps at the crossing time t_s, with with $\lambda_I(t_s^+) - \lambda_I(t_s^-) = \delta_1/(r+\rho+\gamma)$. This positive jump reflects the benefit of accumulating inventories, near the crossing time. Given the jump, one can solve for the time path of $\lambda_I(\,\cdot\,)$, for $t < t_s$.

The second step is to use Corollary 2. It implies that, for any buffer allocation (μ^m, I^m, u^m),

$$\begin{aligned} W(m) - W(0) \;=\; & -\int_0^{+\infty} e^{-rt} w_{hn}(t)\big(\rho\mu^m_{hn}(t) - u^m_h(t)\big)dt \\ & -\int_0^{+\infty} e^{-rt} w_{\ell o}(t)\big(\rho\mu^m_{\ell o}(t) - u^m_\ell(t)\big)dt \\ & -\int_0^{+\infty} e^{-rt}\eta_I(t) I^m(t)dt \\ & +\big(\lambda_I(t_s^+) - \lambda_I(t_s^-)\big)\, I^m(t_s)e^{-rt_s}. \end{aligned} \tag{80}$$

The first two terms in equation (80) are zero because, for all $t \leq t_s$, $u^m_h(t) = \rho\mu^m_{hn}(t)$ and, for all $t \geq t_s$, $u^m_\ell(t) = \rho\mu^m_{\ell o}(t)$. The third term can be bounded as follows

$$0 \leq \int_{\psi(m)-\phi_1(m)}^{\psi(m)+\phi_2(m)} e^{-rt}\eta_I(t) I^m(t)dt \leq \left(\delta_2 - \gamma_d\frac{\delta_1}{r+\rho+\gamma}\right) m\big(\phi_2(m) + \phi_1(m)\big).$$

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

$$\frac{1}{m}\int_{\psi(m)-\phi_1(m)}^{\psi(m)+\phi_2(m)} e^{-rt}\eta_I(t) I^m(t)dt \longrightarrow 0. \tag{81}$$

The fourth and last term of (80) is

$$\big(\lambda_I(t_s^+) - \lambda_I(t_s^-)\big)e^{-rt_s} I^m(t_s). \tag{82}$$

One can make a Taylor expansion of $I^m(t_s)$ around $m = 0$, using the expression (69) for $I^m(t_s)$, the fact that $t_s - t_m = \psi(0) - \psi(m)$, and that $ye^{-\gamma t_m} = y - s + m$. The expansion shows that

$$\frac{1}{m} I^m(t_s) \longrightarrow 1, \tag{83}$$

as m goes to zero, establishing Proposition 2.

B Optimal Allocations

This Appendix solves for constrained-optimal allocations with inventory bound $M \in [0, +\infty]$.

B.1 First-Order Sufficient Conditions

This paragraph provides the first order sufficient conditions of Theorem 3 and Proposition 6. The two identities $\mu_{ho}(t) = \mu_h(t) - \mu_{hn}(t)$ and $\mu_{\ell n}(t) = 1 - \mu_h(t) - \mu_{\ell n}(t)$ have been substituted into the objective and the constraints. The current value Lagrangian is

$$\begin{aligned}\mathcal{L}(t) = \quad & \mu_h(t) - \mu_{hn}(t) + (1-\delta_1)\mu_{\ell o}(t) + (1-\delta_2)I(t) \\ & + \lambda_{\ell o}(t)\big(-u_\ell(t) - \gamma_u\mu_{\ell o}(t) - \gamma_d\mu_{hn}(t) + \gamma_d\mu_h(t)\big) \\ & + \lambda_{hn}(t)\big(-u_h(t) - \gamma_u\mu_{\ell o}(t) - \gamma_d\mu_{hn}(t) + \gamma_u(1-\mu_h(t))\big) \\ & + \lambda_I(t)\big(u_\ell(t) - u_h(t)\big) \\ & + w_{\ell o}(t)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) + w_{hn}(t)\big(\rho\mu_{hn}(t) - u_h(t)\big). \\ & + \eta_I(t)\, I(t) + \eta_M(t)\big(M - I(t)\big).\end{aligned} \tag{84}$$

It is anticipated that the left constraints in (7) and (8) never bind. The first-order conditions are

$$w_{\ell o}(t) = -\lambda_{\ell o}(t) + \lambda_I(t) \tag{85}$$
$$w_{hn}(t) = -\lambda_{hn}(t) - \lambda_I(t) \tag{86}$$
$$r\lambda_{\ell o}(t) = 1 - \delta_1 - \gamma_u(\lambda_{hn}(t) + \lambda_{\ell o}(t)) + \rho w_{\ell o}(t) + \dot{\lambda}_{\ell o}(t) \tag{87}$$
$$r\lambda_{hn}(t) = -1 - \gamma_d(\lambda_{hn}(t) + \lambda_{\ell o}(t)) + \rho w_{hn}(t) + \dot{\lambda}_{hn}(t) \tag{88}$$
$$r\lambda_I(t) = 1 - \delta_2 + \eta_I(t) - \eta_M(t) + \dot{\lambda}_I(t). \tag{89}$$

The complementary-slackness conditions for $w_{\ell o}(t)$, $w_{hn}(t)$, $\eta_I(t)$, and $\eta_M(t)$ are

$$w_{\ell o}(t) \geq 0 \quad \text{and } w_{\ell o}(t)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) = 0 \tag{90}$$
$$w_{hn}(t) \geq 0 \quad \text{and } w_{hn}(t)\big(\rho\mu_{hn}(t) - u_h(t)\big) = 0 \tag{91}$$
$$\eta_I(t) \geq 0 \quad \text{and } \eta_I(t)I(t) = 0 \tag{92}$$
$$\eta_M(t) \geq 0 \quad \text{and } \eta_M(t)\big(M - I(t)\big). \tag{93}$$

Additionally, one imposes the transversality condition

$$\lim_{t\to+\infty} \lambda_j(t)e^{-rt} = 0, \tag{94}$$

for $j \in \{\ell o, hn, I\}$. The multipliers $\lambda_{hn}(t)$ and $\lambda_{\ell o}(t)$ are restricted to be continuous. However, because of the state constraints, the multiplier $\lambda_I(t)$ can jump, with the restrictions

$$\lambda_I(t^+) - \lambda_I(t^-) \leq 0 \quad \text{if } I(t) = 0 \tag{95}$$
$$\lambda_I(t^+) - \lambda_I(t^-) \geq 0 \quad \text{if } I(t) = M. \tag{96}$$

A Reduced System

It is convenient to eliminate $\lambda_{hn}(t)$ and $\lambda_{\ell o}(t)$ from the above equations using the first-order conditions (85) and (86). One obtains the system of ODE

$$\begin{aligned} rw_{\ell o}(t) &= \delta_1 - \delta_2 - \gamma_u\big(w_{hn}(t) + w_{\ell o}(t)\big) - \rho w_{\ell o}(t) \\ &\quad + \eta_I(t) - \eta_M(t) + \dot{w}_{\ell o}(t) \qquad (97) \\ rw_{hn}(t) &= \delta_2 - \gamma_d\big(w_{hn}(t) + w_{\ell o}(t)\big) - \rho w_{hn}(t) \\ &\quad - \eta_I(t) + \eta_M(t) + \dot{w}_{hn}(t) \qquad (98) \\ r\lambda_I(t) &= 1 - \delta_2 + \eta_I(t) - \eta_M(t) + \dot{\lambda}_I(t), \qquad (99) \end{aligned}$$

with the jump conditions

$$\lambda_I(t^+) - \lambda_I(t^-) \leq 0 \text{ if } I(t) = 0 \tag{100}$$
$$\lambda_I(t^+) - \lambda_I(t^-) \geq 0 \text{ if } I(t) = M \tag{101}$$
$$\lambda_I(t^+) - \lambda_I(t^-) = w_{\ell o}(t^+) - w_{\ell o}(t^-) = -w_{hn}(t^+) + w_{hn}(t^-), \tag{102}$$

and the transversality conditions

$$\lim_{t\to+\infty} e^{-rt}\lambda_I(t) = \lim_{t\to+\infty} e^{-rt} w_{\ell o}(t) = \lim_{t\to+\infty} e^{-rt} w_{hn}(t) = 0. \tag{103}$$

As before, the positivity restrictions and complementary slackness conditions are

$$\eta_I(t) \geq 0 \quad \text{and } \eta_I(t)I(t) = 0 \tag{104}$$
$$\eta_M(t) \geq 0 \quad \text{and } \eta_M(t)\big(M - I(t)\big) \tag{105}$$
$$w_{\ell o}(t) \geq 0 \quad \text{and } w_{\ell o}(t)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) = 0 \tag{106}$$
$$w_{hn}(t) \geq 0 \quad \text{and } w_{hn}(t)\big(\rho\mu_{hn}(t) - u_h(t)\big) = 0. \tag{107}$$

Since there is no sign restrictions on the multiplier $\lambda_{hn}(t)$ and $\lambda_{\ell o}(t)$, equations (97)-(107) are equivalent to the first-order sufficient conditions of the previous paragraph.

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 $\big(w_{hn}(t), w_{\ell o}(t), \lambda_I(t), \eta_I(t), \eta_M(t)\big)$ of multipliers solving equations (97)-(107), but ignoring some of the positivity restrictions. These restrictions are imposed afterwards, when discussing the optimality of this allocation. First, one guesses that

$$\eta_M(t) = 0, \tag{108}$$

for all $t \geq 0$. Second, summing equations (97) and (98), and using the transversality condition (103) shows that

$$w_{\ell o}(t) + w_{hn}(t) = \frac{\delta_1}{r + \rho + \gamma}, \tag{109}$$

for all $t \geq 0$. Then, one guesses that there are no jumps at t_1 and t_2.[18] With (102), this shows that

$$\lambda_I(t_i^+) - \lambda_I(t_i^-) = w_{\ell o}(t_i^+) - w_{\ell o}(t_i^-) = -w_{hn}(t_i^+) + w_{hn}(t_i^-) = 0, \tag{110}$$

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 (107) implies that $w_{hn}(t_2) = 0$. With (109), this shows that $w_{\ell o}(t) = \delta_1/(r + \rho + \gamma)$. With (98) and (109), this also implies that $\eta_I(t) = \delta_2 - \gamma_d \delta_1/(r + \rho + \gamma)$. Lastly, (99) and (103) show that $r\lambda_I(t) = 1 - \delta_1 \gamma_d/(r + \rho + \gamma)$.

Time Interval $t \in [t_m, t_2)$

First, because $I(t) > 0$, the complementary slackness condition (104) implies that $\eta_I(t) = 0$. Then, one solves the ODE (98) with the terminal condition $w_{hn}(t_2) = 0$, and one finds that

$$w_{hn}(t) = \frac{1}{r + \rho} \left(\delta_2 - \delta_1 \frac{\gamma_d}{r + \rho + \gamma} \right) \big(1 - e^{(r+\rho)(t - t_2)}\big), \tag{111}$$

[18] This is actually implied by the first-order conditions. Complementary slackness implies that $w_{hn}(t_2^+) = 0$. From (102) and (100), $w_{hn}(t_2^-) \leq 0$. Since $w_{hn}(t_2^-) \geq 0$, this implies that $w_{hn}(t_2^-) = w_{hn}(t_2^+) = 0$. A similar reasoning shows that, if $t_1 > 0$, $w_{\ell o}(t_1^-) = w_{\ell o}(t_1^+) = 0$.

for $t \in [t_m, t_2)$. With (109), $w_{\ell o}(t) = \delta_1/(r+\rho+\gamma) - w_{hn}(t)$. Similarly, one can solve the ODE (99) with the terminal condition $r\lambda_I(t_2^-) = 1 - \gamma_d\delta_1/(r+\rho+\gamma)$, finding that

$$r\lambda_I(t) = 1 - \delta_2 + \left(\delta_2 - \delta_1 \frac{\gamma_d}{r+\rho+\gamma}\right) e^{r(t-t_2)}. \tag{112}$$

Time Interval $t \in [t_1, t_m)$

In this time interval, $\eta_I(t) = 0$. Two cases are considered.

Case 1: $m < \bar{m}$. Complementary slackness at $t = t_1$ shows that $w_{\ell o}(t_1) = 0$, implying that $w_{hn}(t_1) = \delta_1/(r+\rho+\gamma)$. With this and (98), one finds

$$w_{hn}(t) = \frac{1}{r+\rho}\left(\delta_2 - \delta_1\frac{\gamma_d}{r+\rho+\gamma} - \left(\delta_2 - \delta_1\frac{r+\rho+\gamma_d}{r+\rho+\gamma}\right)e^{(r+\rho)(t-t_1)}\right), \tag{113}$$

for $t \in [t_1, t_m)$. Given (111) and (113), the multiplier $w_{hn}(t)$ is not necessarily continuous at time t_m. The size $w_{hn}(t_m^-) - w_{hn}(t_m^+)$ of the jump can be written as some function $b(\,\cdot\,)$ of the maximum inventory level m, where

$$\begin{aligned} b(m) \;\equiv\; & \frac{1}{r+\rho}\Bigg[\left(\delta_2 - \delta_1\frac{\gamma_d}{r+\rho+\gamma}\right)e^{-(r+\rho)\phi_2(m)} \\ & - \left(\delta_2 - \delta_1\frac{r+\rho+\gamma_d}{r+\rho+\gamma}\right)e^{(r+\rho)\phi_1(m)}\Bigg]. \end{aligned} \tag{114}$$

Equation (102) implies that $\lambda_I(t_m^+) - \lambda_I(t_m^-) = b(m)$. This and the ODE (99) show that

$$r\lambda_I(t) = (1-\delta_2) + \big(r\lambda_I(t_m^+) - (1-\delta_2) - rb(m)\big)\, e^{r(t-t_m)}. \tag{115}$$

Case 2: $m = \bar{m}$. Then, by construction of $\bar{m}$, the first breaking time t_1 is equal to zero. If $b(\bar{m}) < 0$, the multipliers are constructed as in the previous case. If, on the other hand, $b(\bar{m}) \geq 0$, then the ODEs are solved so that multipliers *do not* jump at $t = t_m$. Namely, (98) and (99) are solved with terminal conditions $w_{hn}(t_m^-) = w_{hn}(t_m^+)$ and $\lambda_I(t_m^-) = \lambda_I(t_m^+)$. Because $b(\bar{m}) \geq 0$, $w_{hn}(t_1) = w_{hn}(0) \in [0, \delta_1/(r+\rho+\gamma)]$.

Time Interval $t \in [0, t_1]$, $m < \bar{m}$

Complementary slackness shows that $w_{\ell o}(t) = 0$, implying that $w_{hn}(t) = \delta_1/(r+\rho+\gamma)$. With equation (97), this also implies that $\eta_I(t) = \delta_2 - \delta_1(r+\rho+\gamma_d)/(r+\rho+\gamma) \geq 0$. Then

$$r\lambda_I(t) = 1 - \delta_2 - \eta_I(t) + \big(\lambda_I(t_1) - (1-\delta_2) + \eta_I(t)\big)e^{r(t-t_1)}. \tag{116}$$

B.3 Proof of Theorems 1 and Proposition 6

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 (96), 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:

$$m^* = \sup\{m \in [0, \bar{m}] : b(m) \geq 0\}. \tag{117}$$

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

Proposition 8. *There exists a constrained-optimal allocation. It is a buffer allocation with maximum inventory position* $m = \min\{m^*, M\}$.

Proof. Let's consider this allocation and its associated multipliers $\big(w_{hn}(t), w_{\ell o}(t),\ \lambda_I(t), \eta_I(t), \eta_M(t)\big)$, constructed as in the previous subsection. In order to prove optimality, two conditions remain to be verified: the jump conditions (101) and the positivity restrictions in (106) and (107) . Because $m \leq m^*$, the jump condition (101) is satisfied. Also, because $w_{hn}(t)$ is a decreasing function of time, $w_{hn}(0) \in [0, \delta_1/(r + \rho + \gamma)]$ and $w_{hn}(t_2) = 0$, it follows that, at each time, $w_{hn}(t) \in [0, \delta_1/(r + \rho + \gamma)]$, and therefore that $w_{\ell o}(t) \geq 0$.

In particular, the first-order conditions are solved by the multipliers $w_{hn}(t)$, $w_{\ell o}(t)$, $\lambda_I(t)$, $\eta_I(t)$, and $\eta_M(t)$. This proves Proposition 6. The inventory-accumulation period Δ^* 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

$$e^{(r+\rho)\big(\phi_1(m)+\phi_2(m)\big)} \leq 1 + \frac{\delta_1(r + \rho)}{\delta_2\gamma_u + (\delta_2 - \delta_1)(r + \rho + \gamma_d)}. \tag{118}$$

If $m^* < \bar{m}$, then (118) 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 Δ^*, the first breaking time t_1^* is a solution of

$$H(t_1^*, t_1^* + \Delta^*) = 0, \tag{119}$$

and manipulations of (64) give the analytical solution of Theorem 1.

B.4 Proof of Proposition 3

The proof uses the path-comparison result of Corollary 2. The optimal allocation is denoted $\big(\mu^*(t), I^*(t), u^*(t)\big)$. The multipliers associated with this allocation are denoted $\big(w_{hn}(t), w_{\ell o}(t), \lambda_I(t), \eta_I(t)\big)$. These are continuous. Let's consider another allocation $\big(\mu(t), I(t), u(t)\big)$ which achieves the optimum. Corollary 2 implies that

$$\int_0^{+\infty} e^{-rt}\left(w_{hn}(t)\big(\rho\mu_{hn}(t) - u_h(t)\big) + w_{\ell o}(t)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) + \eta_I(t)I(t)\right)dt = 0.$$

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 $u_h(t) = u_\ell(t)$ almost everywhere in $[0, t_1] \cup [t_2, +\infty)$. In $[0, t_1]$, $w_{hn}(t) > 0$ and therefore $u_h(t) = \rho\mu_{hn}(t) = u_\ell(t)$ almost everywhere. In $[t_2, +\infty)$, $w_{\ell o}(t) > 0$ and therefore $u_\ell(t) = \rho\mu_{\ell o}(t) = u_h(t)$ almost everywhere. Lastly, in $[t_1, t_2]$, both $w_{hn}(t) > 0$ and $w_{\ell o}(t) > 0$, implying that $u_h(t) = \rho\mu_{ho}(t)$, and $u_\ell(t) = \rho\mu_{\ell n}(t)$, almost everywhere. Therefore, the allocation $\big(\mu(t), I(t), u(t)\big)$ is equal to $\big(\mu^*(t), I^*(t), u^*(t)\big)$, almost everywhere. This proves Proposition 3.

B.5 Proof of Proposition 4

From Theorem 1, the length of the inventory accumulation period is $\Delta^* = \min\{D(x), \bar{\Delta}\}$, where

$$D(x) = \frac{1}{r+\rho}\log\left(1 + \frac{\delta_1(r+\rho)}{\delta_2\gamma_u + (\delta_2 - \delta_1)(r+\rho+\gamma_d)}\right). \tag{120}$$

The function $D(\,\cdot\,)$ is clearly increasing in δ_1, decreasing in δ_2, γ_d and γ_u. It remains to show that it is also decreasing in r and ρ. Holding fixed $(\delta_1, \delta_2, \gamma_u, \gamma_d)$, and letting $z \equiv r + \rho$, one can write

$$D(x) = \frac{1}{z}\log(1 + \alpha(z)) \equiv \eta(z). \tag{121}$$

where, $\alpha(z) = z/(a + bz)$, for some $(a, b) \in \mathbb{R}^2_+$. The first derivative of $\eta(\,\cdot\,)$ is

$$\frac{d\eta}{dz} = -\frac{1}{z^2}\left(\log\big(1 + \alpha(z)\big) - \frac{z\alpha'(z)}{1 + \alpha(z)}\right) \equiv -\frac{1}{z^2}\beta(z), \tag{122}$$

where $\alpha'(z)$ denotes $d\alpha/dz$. In turn,

$$\frac{d\beta}{dz} = -z\frac{d}{dz}\left(\frac{\alpha'(z)}{1 + \alpha(z)}\right). \tag{123}$$

Since $\alpha'(z)/\big(1 + \alpha(z)\big) = a/(a + bz)/\big(a + (1 + b)z\big)$ is decreasing for $z \geq 0$, it follows that that $d\beta/dz \geq 0$. Since $\beta(0) = 0$, $\beta(z) \geq 0$. Therefore, from (122), $d\eta/dz \leq 0$. This establishes that $D(\,\cdot\,)$ is decreasing in r and ρ.

B.6 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 (68), one can write

$$t_m^* = -\frac{1}{\gamma}\log\left(1 + \frac{s - m^*}{y}\right). \tag{124}$$

Since (124) does not depend on ρ, it follows that $m^* \to 0$ as $\rho \to +\infty$.

C Market Equilibrium

This Appendix proves the various implementation results of the text.

C.1 Proof of Theorem 2 and Proposition 7

The idea of the proof is to identify equilibrium objects with multipliers of constrained-optimal allocations. The current value Lagrangian for the representative market-maker's problem is

$$\begin{aligned}\mathcal{L}(t) \quad &= c(t) + \hat{\lambda}_I(t)\big(u_\ell(t) - u_h(t)\big)\\ &+\hat{\lambda}_a(t)\big(ra(t) + (1-\delta_2)I(t) + p(t)(u_h(t) - u_\ell(t)) - c(t)\big)\\ &+\hat{\eta}_I(t)I(t) + \hat{\eta}_a(t)a(t) + \hat{w}_c(t)c(t),\end{aligned} \tag{125}$$

where δ_2 is set to 1 in the case of Proposition 7. Following Corollary 1, the first-order sufficient conditions are

$$1 + \hat{w}_c(t) = \hat{\lambda}_a(t) \tag{126}$$
$$\hat{\lambda}_I(t) = \hat{\lambda}_a(t)p(t) \tag{127}$$
$$r\hat{\lambda}_I(t) = (1-\delta_2)\hat{\lambda}_a(t) + \hat{\eta}_I(t) + \dot{\hat{\lambda}}_I(t) \tag{128}$$
$$\dot{\hat{\lambda}}_a(t) = -\hat{\eta}_a(t) \tag{129}$$
$$\hat{w}_c(t) \geq 0 \quad \text{and} \quad \hat{w}_c(t)c(t) = 0 \tag{130}$$
$$\hat{\eta}_I(t) \geq 0 \quad \text{and} \quad \hat{\eta}_I(t)I(t) = 0 \tag{131}$$
$$\hat{\eta}_a(t) \geq 0 \quad \text{and} \quad \hat{\eta}_a(t)a(t) = 0 \tag{132}$$
$$\hat{\lambda}_a(t^+) - \hat{\lambda}_a(t^-) \leq 0 \quad \text{if } a(t) = 0, \tag{133}$$

together with the transversality conditions

$$\lim_{t\to+\infty} \lambda_x(t)x(t)e^{-rt} = 0, \tag{134}$$

for $x \in \{I, a\}$. The Bellman equations for the reservation values (see Section D.3) are

$$r\Delta V_\ell(t) = 1 - \delta_1 + \gamma_u\big(\Delta V_h(t) - \Delta V_\ell(t)\big) + \rho\big(p(t) - \Delta V_\ell(t)\big) + \Delta\dot{V}_\ell(t) \quad (135)$$
$$r\Delta V_h(t) = 1 + \gamma_d\big(\Delta V_\ell(t) - \Delta V_h(t)\big) - \rho\big(\Delta V_h(t) - p(t)\big) + \Delta\dot{V}_h(t). \quad (136)$$

The optimality conditions are

$$p(t) - \Delta V_\ell(t) \geq 0 \quad \text{and} \quad \big(p(t) - \Delta V_\ell(t)\big)\big(\rho\mu_{\ell o}(t) - u_\ell(t)\big) = 0. \quad (137)$$
$$\Delta V_h(t) - p(t) \geq 0 \quad \text{and} \quad \big(\Delta V_h(t) - p(t)\big)\big(\rho\mu_{hn}(t) - u_h(t)\big) = 0 \quad (138)$$

and the transversality conditions are

$$\lim_{t\to+\infty} \Delta V_i(t)e^{-rt} = 0, \quad (139)$$

for $i \in \{h, \ell\}$. One proves Theorem 2 and Proposition 7 as follows. One first solves for constrained-optimal allocations with inventory bound $M \in [0, m^*]$. This provides a buffer allocation and a collection $\big(w_{hn}(t), w_{\ell o}(t), \lambda_I(t), \eta_I(t), \eta_M(t)\big)$ of multipliers, solving a system (85)-(95) of first-order sufficient conditions. Then one let $\lambda_{\ell o}(t) = \lambda_I(t) - w_{\ell o}(t)$ and $\lambda_{hn}(t) = -\lambda_I(t) - w_{hn}(t)$. Direct comparison shows that a solution of the system (126)-(139) of equilibrium equations is

$$p(t) = \lambda_I(t) \quad (140)$$
$$\hat{\lambda}_a(t) = 1 + \left(\frac{\lambda_I(t_m^+)}{\lambda_I(t_m^-)} - 1\right)\mathbb{I}_{\{t<t_m\}} \quad (141)$$
$$\hat{\eta}_a(t) = 0 \quad (142)$$
$$\hat{\lambda}_I(t) = \hat{\lambda}_a(t)\lambda_I(t) \quad (143)$$
$$\hat{\eta}_I(t) = \hat{\lambda}_a(t)\eta_I(t) \quad (144)$$
$$\hat{w}_c(t) = \hat{\lambda}_a(t) - 1 \quad (145)$$
$$\Delta V_\ell(t) = \lambda_{\ell o}(t) \quad (146)$$
$$\Delta V_h(t) = -\lambda_{hn}(t), \quad (147)$$

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

$$c(t) = 0 \quad \text{for} \quad t \leq t_2 \quad (148)$$
$$c(t) = ra(t_2) \quad \text{for} \quad t \geq t_2. \quad (149)$$

In this setting, other consumption streams would be optimal. In particular, it is enough that $c(t) = 0$ for $t \in [0, t_m]$ and that $\lim_{t\to+\infty} a(t)e^{-rt} = 0$.

If $M = m^*$ and $\delta_2 \leq 1$, one chooses some large $a(0)$. If $M \leq m^*$ and $\delta_2 = 1$, one lets $a(0) = M\lambda_I(t_m^-)e^{-rt_m}$, which implies that $a(t_m) = 0$, meaning that (133) is satisfied.

To conclude the optimality verification argument for a marketmaker, one needs to check that $a(t) \geq 0$ for all $t \geq 0$. This is equivalent to $a(t)e^{-rt} \geq 0$, for all $t \geq 0$. If $M = m^*$ and $\delta_2 \leq 1$, this is clearly verified provided that $a(0)$ is chosen sufficiently large. If, on the other hand, $M \leq m^*$ and $\delta_2 = 1$, one notes that, for $t \in [t_1, t_2]$, $d/dt\big(a(t)e^{-rt}\big) = -p(t)\dot{I}(t)$ and $\dot{I}(t_m) = 0$. This implies that that $a(t)e^{-rt}$ is continuously differentiable and achieves its minimum at $t = t_m$. By construction $a(t_m) = 0$.

D Investors' Bellman Equations

This Appendix defines the stochastic control problem faced by an individual investor. Then, it verifies that the continuation value equations (49) and (50), the transversality condition (51), together with the positivity restrictions (52) and (53), are jointly sufficient for optimality.

D.1 The Investor's Problem

Some investor is fixed. At each time, his marginal utility for holding asset is some $\theta(t) \in \{1, 1-\delta_1\}$, and he holds a quantity $q(t) \in [0, 1]$ of the asset. Two measurable counting process $N_1(t)$ and $N_2(t)$ count the switching times of the marginal-utility process $\theta(t)$ and the contact times with marketmakers, respectively. The sequence of contact and switching times is denoted by $T_0 = 0 < T_1 < \dots T_n \dots$ The internal history of (filtration generated by) $\big(N_1(t), N_2(t)\big)$ is denoted $\{\mathcal{F}_t^N, t \geq 0\}$.

Definition 6 (Type Process.) *A type process is some $\mathcal{F}_t^N$-adapted, $\{1, 1-\delta_1\} \times [0, 1]$-valued process $\sigma(t) = \big(\theta(t), q(t)\big)$. An admissible control is some $\mathcal{F}_t^N$-predictable, $[0, 1]$-valued process $Q(t)$. The set of admissible controls is denoted by $\mathcal{Q}$.*

Given some control $Q \in \mathcal{Q}$, the type process $\sigma(t) = \big(\theta(t), q(t)\big)$ evolves according to the stochastic differential equation (SDE)

$$d\theta(t) = \big(2 - \delta_1 - 2\theta(t^-)\big)\, dN_1(t) \tag{150}$$

$$dq(t) = \big(Q(t) - q(t^-)\big)\, dN_2(t), \tag{151}$$

meaning that the investor switches type when $dN_1(t) = 1$, and rebalances his holding to the quantity $Q(t)$ when $dN_2(t) = 1$. The associated cumulative consumption process evolves according to the SDE

$$dC^Q(t) = \theta(t)q(t)\, dt - p(t)\, dq(t), \tag{152}$$

where, at time t, the investor buys or sells the asset at price $p(t)$. Lastly, the associated probability P on $(\Omega, \mathcal{F})$ is chosen such that $\big(N_1(t), N_2(t)\big)$ admits the $P - \mathcal{F}^N$ intensity $\big(\gamma(\theta(t^-)), \rho\big)$, where $\gamma(1) = \gamma_d$ and $\gamma(1-\delta_1) = \gamma_u$.

Definition 7 (Investor's Problem.) *The lifetime utility of an investor applying the admissible control $Q \in \mathcal{Q}$ is*

$$v(Q) = E_P\left(\int_0^{+\infty} e^{-rt} dC^Q(t)\right). \tag{153}$$

The investor's problem is to attain the maximum lifetime utility

$$V = \sup_{Q \in \mathcal{Q}} v(Q). \tag{154}$$

D.2 Dynamic Programming

The investor's Hamilton-Jacobi-Bellman (HJB) equation is

$$\begin{aligned} rJ(t,\theta,q) = \max_{Q\in[0,1]} \Big\{ & \theta q + \gamma(\theta)\big(J(t, 2-\delta_1-\theta, q) - J(t,\theta,q)\big) \\ & + \rho\big(J(t,\theta,Q) - J(t,\theta,q) - p(t)(Q-q)\big) + \frac{\partial J}{\partial t}(t,\theta,q)\Big\}. \end{aligned} \tag{155}$$

Definition 8 (Admissible Feedback.) *An admissible feedback is some function $F : \mathbb{R}_+ \times \{1, 1-\delta_1\} \times [0,1] \to [0,1]$.*

The following proposition, adapted from Theorem VII, T1 in Brémaud [1981], provides a sufficient condition for optimality.

Proposition 9 (Sufficiency.) *Suppose there exists a function $J(t,\theta,q)$ which is bounded, continuous and piecewise continuously differentiable with respect to time, and which solves the HJB equation (155). Then,*

$$V \leq J\big(0,\, \theta(0),\, q(0)\big), \tag{156}$$

for all $\big(\theta(0), q(0)\big) \in \{1, 1-\delta_1\} \times [0,1]$. Suppose further that, given $J(t,\theta,q)$, the maximum in (155) is achieved by some admissible feedback $F(t,\theta,q)$. Then, the investor's problem is solved by the admissible control Q such that, at each time

$$Q(t) = F\big(t,\, \theta(t^-),\, q(t^-)\big). \tag{157}$$

Proof. One fixes some admissible control $Q \in \mathcal{Q}$. The associated type process is denoted $\sigma(t)$. To simplify the notations, $J(t,\theta,q)$ is denoted $J(t,\sigma)$, and the following predictable processes are defined

$$\sigma_1(t) \equiv \big(2-\delta_1-\theta(t^-),q(t^-)\big) \tag{158}$$
$$\sigma_2(t) \equiv \big(\theta(t^-),Q(t)\big). \tag{159}$$

The process $\sigma_1(t)$ ($\sigma_2(t)$) is the new type of an investor with control $Q(t)$ who, at time t, switches marginal utility (establishes contact with a marketmaker). One can write:

$$\begin{aligned} e^{-rt}J\big(t,\sigma(t)\big) &= J\big(0,\sigma(0)\big) \\ &\quad + \sum_{0<T_n\le t}\Big(e^{-rT_n}J\big(T_n,\sigma(T_n)\big)-e^{-rT_{n-1}}J\big(T_{n-1},\sigma(T_{n-1})\big)\Big) \\ &\quad + e^{-rt}J\big(t,\sigma(\tau_t)\big)-e^{-r\tau_t}J\big(\tau_t,\sigma(\tau_t)\big), \end{aligned} \tag{160}$$

where $\tau_t=\sup\{T_n,n\ge 0:T_n\le t\}$. Equation (160) can be manipulated as follows:

$$\begin{aligned} e^{-rt}J\big(t,\sigma(t)\big) &= J\big(0,\sigma(0)\big)+\sum_{0<T_n\le t}e^{-rT_n}\Big(J\big(T_n,\sigma(T_n)\big)-J\big(T_n,\sigma(T_{n-1})\big)\Big) \\ &\quad + \sum_{0<T_n\le t}\Big(e^{-rT_n}J\big(T_n,\sigma(T_{n-1})\big)-e^{-rT_{n-1}}J\big(T_{n-1},\sigma(T_{n-1})\big) \\ &\quad + \Big(e^{-rt}J\big(t,\sigma(\tau_t)\big)-e^{-r\tau_t}J\big(\tau_t,\sigma(\tau_t)\big)\Big). \end{aligned}$$

The second term on the right-hand side collects jumps of the value function at switching and contact times. These can be rewritten using the two predictable processes $\sigma_1(t)$ and $\sigma_2(t)$. The third and the fourth terms collect the time variation of the value function between switching and contact times, when the type process stays constant. Since the value function is continuous and piecewise continuously differentiable with respect to time, these can be written as the integral of $\partial/\partial t\big(e^{-rt}J(t)\big)$. This implies that

$$
\begin{aligned}
e^{-rt}J\big(t,\sigma(t)\big) &= J(0) + \int_0^t e^{-rz}\Big(J\big(z,\sigma_1(z)\big) - J\big(z,\sigma(z^-)\big)\Big)\,dN_1(z) \\
&\quad + \int_0^t e^{-rz}\Big(J\big(z,\sigma_2(z)\big) - J\big(z,\sigma(z^-)\big)\Big)\,dN_2(z) \\
&\quad + \int_0^t \frac{\partial}{\partial t}\Big(e^{-rz}J\big(z,\sigma(z)\big)\Big)\,dz \\
&= J(0) + \int_0^t e^{-rz}\Big(J\big(z,\sigma_1(z)\big) - J\big(z,\sigma(z^-)\big)\Big)\big(dN_1(z) - \gamma(\theta(z^-))\,dz\big) \\
&\quad + \int_0^t e^{-rz}\Big(J\big(z,\sigma_2(z)\big) - J\big(z,\sigma(z^-)\big)\Big)\big(dN_2(z) - \rho\,dz\big) \\
&\quad + \int_0^t \Big[-rJ\big(z,\sigma(z)\big) + \frac{\partial J}{\partial t}\big(z,\sigma(z)\big) + \gamma(\theta(z^-))\Big(J\big(z,\sigma_2(z)\big) - J\big(z,\sigma(z^-)\big)\Big) \\
&\qquad\qquad + \rho\Big(J\big(z,\sigma_1(z)\big) - J\big(z,\sigma(z^-)\big)\Big)\Big]\,dz, \qquad (161)
\end{aligned}
$$

Adding $\int_0^t e^{-rz}dC^Q(z)$ to both sides gives

$$
\begin{aligned}
&\int_0^t e^{-rz}dC^Q(z) + e^{-rt}J(t) \qquad (162) \\
= \;& J(0) + \int_0^t e^{-rz}\Big(J\big(z,\sigma_1(z)\big) - J\big(z,\sigma(z^-)\big)\Big)\big(dN_1(z) - \gamma(\theta(z^-))\,dz\big) \\
&+ \int_0^t e^{-rz}\Big(J\big(z,\sigma_2(z)\big) - J\big(z,\sigma(z^-)\big)\Big)\big(dN_2(z) - \rho\,dz\big) \\
&- \int_0^t e^{-rz}p(z)\big(Q(z) - q(z^-)\big)\big(dN_1(z) - \rho\,dz\big) \\
&+ \int_0^t \Big[-rJ\big(z,\sigma(z)\big) + \frac{\partial J}{\partial t}\big(z,\sigma(z)\big) + \gamma(\theta(z^-))\Big(J\big(z,\sigma_1(z)\big) - J\big(z,\sigma(z^-)\big)\Big) \\
&\qquad + q(z)\theta(z) + \rho\Big(J\big(z,\sigma_2(z)\big) - J\big(z,\sigma(z^-)\big) - p(z)\big(Q(z) - q(z^-)\big)\Big)\Big]\,dz.
\end{aligned}
$$

Because J is bounded, and because $\sigma_1(z)$, $\sigma_2(z)$, $\sigma(z^-)$, $Q(z)$, and $q(z^-)$ are $\mathcal{F}_t^N$-predictable processes, it follows by Theorem II, T8 in Brémaud [1981] that the first three integrals on the right-hand side of (162) are martingale. The last integral on the right-hand side of (162) is negative because $J(\,\cdot\,)$ solves the Bellman equation. Taking expectations on both sides gives

$$
E_P\left(\int_0^t e^{-rz}dC^Q(z) + J(t)e^{-rt}\right) \le J(0), \qquad (163)
$$

with equality for $Q(t) = F\big(t,\theta(t^-),q(t^-)\big)$. Letting t go to infinity shows that $v(Q) \le J(0)$, with equality for $Q(t) = F\big(t,\theta(t^-),q(t^-)\big)$. This proves the proposition.

D.3 Reservation Values and Value Function

This subsection constructs a solution of the HJB equation (155) using the reservation values $\Delta V_\ell(t)$ and $\Delta V_h(t)$ of investors. These solve

$$\begin{aligned} r\Delta V_\ell(t) &= 1-\delta_1+\gamma_u\big(\Delta V_h(t)-\Delta V_\ell(t)\big)+\rho\big(p(t)-\Delta V_\ell(t)\big)+\Delta\dot{V}_\ell(t), && (164)\\ r\Delta V_h(t) &= 1+\gamma_d\big(\Delta V_\ell(t)-\Delta V_h(t)\big)-\rho\big(\Delta V_h(t)-p(t)\big)+\Delta\dot{V}_h(t), && (165)\end{aligned}$$

and are assumed to satisfy the transversality condition

$$\lim_{T\to+\infty} e^{-rT}\Delta V_j(T)=0, \tag{166}$$

for $j\in\{\ell,h\}$, as well as the positivity restrictions

$$\Delta V_\ell(t)\le p(t)\le \Delta V_h(t). \tag{167}$$

Given these, one solves the ODE

$$\begin{aligned} rV_{\ell n}(t) &= \gamma_u\big(V_{hn}(t)-V_{\ell n}(t)\big)+\dot{V}_{\ell n}(t) && (168)\\ rV_{hn}(t) &= \gamma_d\big(V_{\ell n}(t)-V_{hn}(t)\big)+\rho\big(\Delta V_h(t)-p(t)\big)+\dot{V}_{hn}(t). && (169)\end{aligned}$$

Subtracting (168) from (169), integrating, and using the transversality condition, I find that

$$V_{hn}(t)-V_{\ell n}(t)=\int_t^{+\infty} e^{-(r+\gamma)(z-t)}\rho\big(\Delta V_h(z)-p(z)\big)\,dz. \tag{170}$$

And, replacing (170) in (168), that

$$V_{\ell n}(t)=\int_t^{+\infty} e^{-r(z-t)}\gamma_u\big(V_{hn}(t)-V_{\ell n}(t)\big)dz. \tag{171}$$

Then, it is easy to check that the HJB equation is solved by the value function

$$\begin{aligned} J(t,1,q) &= V_{hn}(t)+q\Delta V_h(t) && (172)\\ J(t,1-\delta_1,q) &= V_{\ell n}(t)+q\Delta V_\ell(t). && (173)\end{aligned}$$

with some feedback F such that $F(t,1,q)=1$ if $p(t)<\Delta V_h(t)$, $F(t,1,q)\in[0,1]$ if $p(t)=\Delta V_h(t)$, $F(t,1-\delta_1,q)=0$ if $p(t)>\Delta V_\ell(t)$, and $F(t,1-\delta_1,q)\in[0,1]$ if $p(t)=\Delta V_\ell(t)$.

E Maximum Principle

This Appendix defines a concave optimal control problem and provides sufficient condition for optimality. The problem includes as special cases all the planner's and marketmakers' problems considered in the text. Two features require some care: first, there are state-variable inequality constraints and second, at some Poisson arrival time, there is an uncontrollable jump of the state variable (this can be used to treat the case of a fully anticipated crash). The proofs use standard optimality-verification arguments from Seierstad and Sydsæter [1977], Brémaud [1981], and Kamien and Schwartz [1991].

E.1 An Optimal Control Problem

Some probability space $(\Omega, \mathcal{F}, P)$ is fixed, as well as some counting process $N(t)$, with initial condition $N(0) = 0$ and admitting the P-intensity

$$\kappa(t) = \kappa \mathbb{I}_{\{N(t)=0\}}, \tag{174}$$

for some $\kappa \geq 0$. In other words, the process $N(t)$ jumps only once (almost surely), at some exponentially distributed stopping time τ. The internal history of (filtration generated by) $N(t)$ is denoted $\{\mathcal{F}_t^N, t \geq 0\}$. Lastly, $y(t) \in \mathbb{R}^{n_y}$ is some bounded $\mathcal{F}_t^N$-adapted piecewise continuous process (this process can be, for instance, a candidate equilibrium price). At each time, the economy is described by some state variable $x(t) \in \mathbb{R}^{n_x}$, with initial condition

$$x(0) = x_0, \tag{175}$$

and evolving according to the ODE

$$\dot{x}(t) = g\big(x(t), u(t), y(t)\big), \tag{176}$$

where $u(t) \in \mathbb{R}^{n_u}$ is some control process, and $g : \mathbb{R}^{n_x+n_u+n_y} \to \mathbb{R}^{n_x}$ is some continuously differentiable function. The state and the control must satisfy the mixed constraint

$$h\big(x(t), u(t), y(t)\big) \geq 0, \tag{177}$$

where $h : \mathbb{R}^{n_x+n_u+n_y} \to \mathbb{R}^m$ is some continuously differentiable function. The state must satisfy the constraint

$$k\big(x(t), y(t)\big) \geq 0, \tag{178}$$

where $k : \mathbb{R}^{n_x+n_y} \to \mathbb{R}^s$ is some continuously differentiable function. Lastly, there is an uncontrollable jump of the state at time τ,

$$x(\tau^+) = Q(x(\tau^-), y(\tau^-)), \tag{179}$$

where $Q : \mathbb{R}^{n_x+n_y} \to \mathbb{R}^{n_x}$ is some continuously differentiable function.

Definition 9 (State-Control Pair) . *A state-control pair is some $\mathcal{F}_t^N$-adapted piecewise continuous $\mathbb{R}^{n_x+n_u}$-valued process $\big(x(t), u(t)\big)$. A state-control pair is feasible if it satisfies the constraints (175) to (179).*

The dynamic optimization problem is to choose some feasible state-control pair (x, u) in order to maximize the objective

$$E_0\left(\int_0^{+\infty} e^{-rt} f\big(x(t), u(t), y(t)\big)\, dt\right), \tag{180}$$

where $f : \mathbb{R}^{n_x+n_u+n_y} \to \mathbb{R}^m$ is some continuously differentiable function.

E.2 Sufficient Conditions

In order to simplify the exposition, the following notations are adopted. Let $z(t)$ and $z^*(t)$ be two $\mathbb{R}^{n_z}$-valued processes, and let $\theta : \mathbb{R}^{n_z} \to \mathbb{R}^{m_z}$. Then, $\theta(t)$ denotes $\theta\big(z(t)\big)$, and $\theta^*(t)$ denotes $\theta\big(z^*(t)\big)$. Also, the time index is omitted whenever there is no ambiguity. For two vectors z and z^* in $\mathbb{R}^{n_z}$, $z \cdot z^*$ denotes the inner product $\sum_{i=1}^{n_z} z_i z_i^*$. Lastly, the current-value Lagrangian is

$$\mathcal{L}\big(x, u, y\big) = f(x, u, y) + \lambda \cdot g(x, u, y) + w \cdot h(x, u, y) + \eta \cdot k(x, y), \tag{181}$$

where $x \in \mathbb{R}^{n_x}$, $u \in \mathbb{R}^{n_u}$, $y \in \mathbb{R}^{n_y}$, $\lambda \in \mathbb{R}^{n_x}$, $w \in \mathbb{R}^m$, and $\eta \in \mathbb{R}^s$.

Theorem 3 (Sufficient Condition.) *Let $\big(x^*(t), u^*(t)\big)$ be a feasible state-control pair. If all feasible $x(t)$ are bounded. If f, g, h, k and Q are concave in (x, u). If there exists a piecewise continuously differentiable multiplier process $\lambda(t) \in \mathbb{R}^{n_x}$, some piecewise continuous multiplier processes $w(t) \in \mathbb{R}^m$, $\eta(t) \in \mathbb{R}^s$, and some piecewise continuous predictable processes $\Delta(t) \in \mathbb{R}^{n_x}$ and $b(t) \in \mathbb{R}^s$ such that*

(i) Maximization: At each continuity point of $u^*(t)$ and $\lambda(t)$, for all $j \in \{1, \dots n_u\}$

$$0 = \frac{\partial \mathcal{L}^*}{\partial u_j}(t). \tag{182}$$

(ii) Multipliers: At each continuity point of $u^*(t)$ and $\lambda(t)$, for all $i \in \{1, \ldots, n_x\}$,

$$r\lambda_i(t) = \frac{\partial \mathcal{L}^*}{\partial x_i}(t) + \kappa(t)\Delta_i(t) + \dot{\lambda}_i(t), \tag{183}$$

and $\lambda_i(t) \geq 0$ except if, for all $y \in \mathbb{R}^{n_y}$, the function $g_i(x, u, y)$ is affine in (x, u).

(iii) Jumps: At any discontinuity point t of $\lambda(t)$, for all $i \in \{1, \ldots, n_x\}$,

$$t \neq \tau \quad \Rightarrow \quad \lambda_i(t^+) - \lambda_i(t^-) = -\sum_{l=1}^{s} b_l(t) \frac{\partial k_l^*}{\partial x_i}(t) \tag{184}$$

$$t = \tau \quad \Rightarrow \quad \sum_{p=1}^{n_x} \lambda_p(t^+) \frac{\partial Q_p^*}{\partial x_i}(t^-) - \lambda_i(t^-) = \Delta_i(t) \tag{185}$$

and, for all $p \in \{1, \ldots, n_x\}$, $\lambda_p(t^+) \geq 0$ except if, for all $y \in \mathbb{R}^{n_y}$, the function $Q_p(x, y)$ is affine in x.

(iv) Transversality: for all $i \in \{1, \ldots, n_x\}$,

$$\lim_{t \to +\infty} E_0\big(\lambda_i(t) e^{-rt}\big) = 0. \tag{186}$$

(v) Positivity and Complementary Slackness

$$w_p(t) \geq 0 \quad \text{and} \quad w_p(t) h_p^*(t) = 0 \tag{187}$$

$$\eta_l(t) \geq 0 \quad \text{and} \quad \eta_l(t) k_l^*(t) = 0 \tag{188}$$

$$b_l(t) \geq 0 \quad \text{and} \quad b_l(t) k_l^*(t) = 0 \tag{189}$$

for all $p \in \{1, \ldots, m\}$ and all $l \in \{1, \ldots, s\}$.

Then $\big(x^*(t), u^*(t)\big)$ is optimal.

E.3 Proof of Theorem 3

Let $\big(x^*(t), u^*(t)\big)$ be some state-control pair satisfying the condition of the Theorem, and let $\big(x(t), u(t)\big)$ be some other feasible state-control pair. Having fixed some $T \geq 0$, $t_1 < t_2 < \ldots < t_{K-1}$ denotes the discontinuity points of $\big(x^*(t), u^*(t), \lambda(t), x(t), u(t)\big)$ in the interval $[0, T]$, $t_0 \equiv 0$ and $t_K \equiv T$. One considers

$$\begin{aligned} \int_0^T e^{-rt}\big(f^*(t) - f(t)\big)\, dt &= \sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} e^{-rt}\big(f^*(t) - f(t)\big) dt \\ \geq \quad \sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} e^{-rt} &\left(\sum_{i=1}^{n_x} \frac{\partial f^*}{\partial x_i}(x_i^* - x_i) + \sum_{j=1}^{n_u} \frac{\partial f^*}{\partial u_j}(u_j^* - u_j) \right) dt, \end{aligned} \tag{190}$$

where the inequality follows from the concavity of f. Then, with equation (182) and (183), the partial derivatives of f can be written, for all $j \in \{1, \dots, n_u\}$ and all $i \in \{1, \dots, n_x\}$,

$$\begin{aligned}
\frac{\partial f^*}{\partial u_j}(t) &= -\lambda(t) \cdot \frac{\partial g^*}{\partial u_j}(t) - w(t) \cdot \frac{\partial h^*}{\partial u_j}(t) && (191)\\
\frac{\partial f^*}{\partial x_i}(t) &= -\dot{\lambda}_i(t) + r\lambda_i(t) && (192)\\
&\quad -\lambda(t) \cdot \frac{\partial g^*}{\partial x_i}(t) - w(t) \cdot \frac{\partial h^*}{\partial x_i}(t) - \eta(t) \cdot \frac{\partial k^*}{\partial x_i}(t) - \kappa(t)\Delta_i(t),
\end{aligned}$$

Furthermore, integration by part shows that

$$\begin{aligned}
&\int_{t_{k-1}}^{t_k} e^{-rt} \left(-\dot{\lambda}_i(t) + r\lambda_i(t)\right) \left(x_i^*(t) - x_i(t)\right) dt \\
= &\int_{t_{k-1}}^{t_k} e^{-rt} \lambda_i(t) \big(g_i^*(t) - g_i(t)\big)\, dt - \left[\lambda_i(t) e^{-rt} \left(x_i^*(t) - x_i(t)\right)\right]_{t_{k-1}}^{t_k}, \qquad (193)
\end{aligned}$$

for all $k \in \{1, \dots K\}$ and all $i \in \{1, \dots, n_x\}$. Substituting (191), (192) and (193), in equation (190), one finds that (190) is the sum of six terms. The first term is

$$\int_0^T e^{-rt} \lambda(t) \cdot \left(g^*(t) - g(t) - \sum_{i=1}^{n_x} \frac{\partial g^*}{\partial x_i}(x_i^* - x_i) + \sum_{j=1}^{n_u} \frac{\partial g^*}{\partial u_j}(u_j^* - u_j) \right) dt. \qquad (194)$$

This term is positive if either one of the following conditions are satisfied: if $g(\,\cdot\,)$ is concave and λ positive or if, for all y, $g(x, u, y)$ is affine in (x, u), in which case it is equal to zero. The second term is

$$\begin{aligned}
&-\int_0^T e^{-rt} w(t) \cdot \left(\sum_{i=1}^{n_x} \frac{\partial h^*}{\partial x_i}(x_i^* - x_i) + \sum_{j=1}^{n_u} \frac{\partial h^*}{\partial u_j}(u_j^* - u_j) \right) dt \\
\geq\ &-\int_0^T e^{-rt} w(t) \cdot \big(h^*(t) - h(t)\big)\, dt && (195)\\
\geq\ &\int_0^T e^{-rt} w(t) \cdot h(t)\, dt \geq 0 && (196)
\end{aligned}$$

where the first inequality (195) follows from the concavity of h and the positivity of $w(t)$. The second inequality (196) follows from the complementary-slackness condition (187). Similarly, the third term is

$$
\begin{aligned}
-\int_0^T e^{-rt}\eta(t)\cdot\sum_{i=1}^{n_x}\frac{\partial k^*}{\partial x_i}(x_i^*-x_i)\,dt &\geq -\int_0^T e^{-rt}\eta(t)\cdot\big(k^*(t)-\hat{k}(t)\big)\,dt \\
&\geq \int_0^T e^{-rt}\eta(t)\cdot k(t)\,dt \geq 0. \qquad (197)
\end{aligned}
$$

As before, the first inequality follows from the concavity of k and the positivity of $\eta(t)$. The second inequality (197) follows from the complementary-slackness condition (188). The fourth term collects the jumps of $\lambda(t)$ at discontinuity points $t \neq \tau$

$$
\begin{aligned}
&\sum_{t_k\neq\tau} e^{-rt_k}\big(\lambda(t_k^+)-\lambda(t_k^-)\big)\cdot\big(x^*(t_k)-x(t_k)\big) \qquad (198)\\
=\;& -\sum_{t_k\neq\tau} e^{-rt_k}b(t_k)\cdot\sum_{i=1}^{n_x}\frac{\partial k^*}{\partial x_i}(x_i^*-x_i)\\
\geq\;& -\sum_{t_k\neq\tau} e^{-rt_k}b(t_k)\cdot\big(k^*(t_k)-k(t_k)\big)\geq\sum_{t_k\neq\tau} b(t_k)\cdot k(t_k)e^{-rt}dt\geq 0.
\end{aligned}
$$

The first equality follows from the jump condition (184). The second inequality follows from the concavity of k and the positivity of b. And the third inequality follows from the complementary-slackness condition (189).

The fifth term collects the jump of $\lambda(t)$ at τ and the last term in (192). It can be written

$$
\begin{aligned}
&\int_0^T e^{-rt}\Big(\lambda(t^+)\cdot\big(x^*(t^+)-x(t^+)\big)-\lambda(t^-)\cdot\big(x^*(t^-)-x(t^-)\big)\Big)\,dN(t)\\
-\;&\int_0^T e^{-rt}\kappa(t)\Delta(t)\cdot\big(x^*(t)-x(t)\big)\,dt. \qquad (199)
\end{aligned}
$$

The term (199) can be manipulated as follows. First, one uses (179) and the concavity of Q and the positivity of $\lambda(t^+)$ (or, alternatively, the fact that Q is affine in x) to write

$$
\begin{aligned}
&\lambda(t^+)\cdot\big(x^*(t^+)-x(t^+)\big)-\lambda(t^-)\cdot\big(x^*(t^-)-x(t^-)\big)\\
\geq\;&\left(\lambda(t^+)\sum_{i=1}^{n_x}\frac{\partial Q^*}{\partial x_i}(t^-)-\lambda(t^-)\right)\cdot\big(x^*(t^-)-x(t^-)\big). \qquad (200)
\end{aligned}
$$

Then, one substitutes the jump conditions (185) and collect terms to obtain that (199) is greater than

$$\int_0^T e^{-rt}\Delta(t)\cdot\big(x^*(t^-)-x(t^-)\big)\big(dN(t)-\kappa(t)dt\big). \tag{201}$$

Since $\Delta(t)\cdot\big(x^*(t^-)-x(t^-)\big)$ is a bounded $\mathcal{F}_t^N$-predictable process on $[0,T]$, it follows by theorem II, T8 in Brémaud [1981] that (201) is a martingale. This implies that the expected value of (199) is greater than zero. The last and sixth term is

$$-e^{-rT}\lambda(T)\cdot\big(x^*(T)-x(T)\big). \tag{202}$$

Collecting the six terms just studied and taking expectations, one finds

$$E_0\left(\int_0^T e^{-rt}\big(f^*(t)-f(t)\big)dt\right)\geq -E_0\left(e^{-rT}\lambda(T)\cdot\big(x^*(T)-x(T)\big)\right). \tag{203}$$

The assumption that all feasible states are bounded, and the transversality condition (186) imply that

$$E_0\left(\int_0^{+\infty} e^{-rt}\big(f^*(t)-f(t)\big)dt\right)\geq 0. \tag{204}$$

establishing the Theorem.

E.4 Two Corollaries

Theorem 3 assumed that all feasible states are bounded, which is enough to prove optimality in all planner's problems under consideration. Proving optimality in Section 5 requires, however, alternative assumptions

Corollary 1 . *Let $\big(x^*(t),u^*(t)\big)$ be a feasible state-control pair. Assume that all assumptions of Theorem 3 are satisfied except the first. Namely, instead of assuming that all feasible states are bounded, assume that there exists some $B\in\mathbb{R}^{n_x}$ such that, for all $i\in\{1,\ldots,n_x\}$*

(i) $\lim_{T\to+\infty} e^{-rT}\lambda_i(T)\big(x_i^*(T)-B_i\big)=0$, *almost surely.*

(ii) For all feasible states $x(t)$,

$$\text{either}\quad x_i(t)\leq B_i\quad\text{and}\quad \limsup_{T\to+\infty}\lambda_i(T)<0,\ \text{a.s.} \tag{205}$$

$$\text{or}\quad x_i(t)\geq B_i\quad\text{and}\quad \liminf_{T\to+\infty}\lambda_i(T)>0,\ \text{a.s.} \tag{206}$$

Then $\big(x^*(t), u^*(t)\big)$ *is optimal*

Proof. The same proof can be applied, with one slight difference: term (202) has to be written

$$\sum_{i=1}^{n_x} e^{-rT}\bigg(- \lambda_i(T)\big(x_i^*(T) - B_i\big) + \lambda_i(T)\big(x_i(T) - B_i\big)\bigg). \tag{207}$$

The first term goes to zero almost surely, and the second term is positive for T large enough, implying (204).

In the proof of Theorem 3, all inequalities implied by concavity become equalities when the functions f, g, h, k, and Q are affine in (x, u). This remark implies the following "path-comparison" corollary. Specifically, given a feasible state-control pair $\big(x^*(t), u^*(t)\big)$, one constructs multipliers satisfying the conditions of Theorem 3, except the positivity restrictions and the jump conditions (184). These multipliers can be used to compare the value of the objective at $\big(x^*(t), u^*(t)\big)$ with its value at some other feasible pair $\big(x(t), u(t)\big)$.

Corollary 2 (Path Comparison.) . *Let* $\big(x^*(t), u^*(t)\big)$ *and* $\big(x(t), u(t)\big)$ *be two feasible state-control pairs. If both* $x^*(t)$ *and* $x(t)$ *are bounded. If* f, g, h, k *and* Q *are affine in* (x, u). *If there exists multiplier processes* $(\lambda(t), w(t), \eta(t))$ *satisfying all conditions of Theorem 3 except the jump condition (184) and the positivity of* $(w(t), \eta(t))$. *Then*

$$\begin{aligned} & E_0\left(\int_0^{+\infty} e^{-rt}\left(f^*(t) - f(t)\right) dt\right) \\ = \quad & E_0\left(\int_0^{+\infty} e^{-rt}\big(w(t)h(t) + \eta(t)k(t)\big)dt\right) \\ + \quad & E_0\left(\sum_{t_k \neq \tau} \big(x^*(t_k) - x(t_k)\big) \cdot \big(\lambda(t_k^+) - \lambda(t_k^-)\big)e^{-rt_k}\right), \end{aligned} \tag{208}$$

where $t_1, t_2, \ldots$ *are the discontinuity points of* $\lambda(t)$.

F Capital Market Imperfection

This Appendix extends the model of this paper by studying a capital market imperfection associated with moral hazard. The key impact of such an imperfection is to increase the cost of lending to marketmakers. In equilibrium, in order to cover these costs, marketmakers must make positive profit. This implies in turns that the aggregate quantity of marketmaking capital must be less than optimal. In the environment presented here, welfare can be improved by subsidizing the loans made to marketmakers. During financial disruptions, Federal Reserves take actions which

effectively subsidize bank lending. For instance, they offer cheap discount window borrowing and relax overdraft penalties (see, for instance, Parry [1997] or Wigmore [1998]).

The Appendix is organized as follows. The first Section introduces some notations. The second Section studies an environment with moral hazard, and the third an environment with adverse selection.

F.1 The Economic Environment

Let us consider the following extension of the model presented in Section 5. There is a unit measure of competitive banks (owned by investors), each endowed with a quantity $\bar{x}$ of capital. Marketmakers, on the other hand, are wealthless and have limited-liability. At time zero, when the crash occurs, banks and marketmakers are matched in pairs. The bank makes a take-it-or-leave-it offer to the marketmaker, consisting in a loan size $a \in [0, \bar{x}]$, and a repayment $R\,a$ (in present value), to be made after the crash. Suppose that the marketmaker accepts the contract. The analysis of Section 5 shows that, in equilibrium, the marketmaker's optimal strategy is to spend all her capital between some time t_1 and some time t_m.[19] At time t_m, the asset price jumps from $p(t_m^-)$ to $p(t_m^+) \geq p(t_m^-)$. As a result of this upward jump, the marketmaker intertemporal profit is

$$\left(\frac{p(t_m^+)}{p(t_m^-)} - R\right) a. \tag{209}$$

The size $p(t_m^+)/p(t_m^-)$ of the jump is implicitly a function of the aggregate marketmaking capital $x \in [0, \bar{x}]$. In this Appendix, I make this dependence explicit and write (209) as

$$\big(1 + F(x) - R\big)\, a, \tag{210}$$

for some positive function $F(\,\cdot\,)$ of the aggregate marketmaking capital. I refer to $F(x)$ as the return on marketmaking capital. Of course, a bank profit is

$$(R-1)\, a + \bar{x}. \tag{211}$$

The analysis of Section 5 also shows that equilibrium social welfare is $W_e(x) + \bar{x}$, for some function of the aggregate marketmaking capital $x \in [0, \bar{x}]$. The analysis of this Appendix requires the following result, shown in Section F.3:

Lemma 3. *If $\bar{x}$ is small enough, then the function W_e (respectively F) is strictly increasing (decreasing) on the interval $[0, \bar{x}]$.*

[19] The precise timing is irrelevant as long as all the capital is spent during this time interval.

It is also known that, if x is large enough, then $W_e(x)$ is maximized and $F(x)$ is equal to zero. Numerical calculations suggest that, for $x \in \mathbb{R}_+$, these two functions are weakly increasing and decreasing, respectively.

Perfect Capital Market

Since $W_e(x)$ is an increasing function, the socially-optimal allocation of capital is that marketmakers receive all the capital at time zero. This allocation is implemented by the trading arrangement described in the previous Section. The bank choose a loan size $a \in [0, \bar{x}]$ and a repayment R in order to maximize profit (211), subject to the marketmaker's limited liability

$$1 + F(x) - R \geq 0. \tag{212}$$

Clearly, it is optimal for the bank to offer $R = 1 + F(x)$, which is the largest repayment consistent with limited liability. Substituting this into the bank's objective, one defines an equilibrium as some $x^* \in [0, \bar{x}]$ such that

$$x^* \in \underset{a \in [0,\bar{x}]}{\operatorname{argmax}} \left\{ a\, F(x^*) + \bar{x} \right\}. \tag{213}$$

Clearly, $x^* = \bar{x}$ is an equilibrium.

F.2 Moral Hazard

This subsection studies a simple capital market imperfection based on moral hazard. It implies that subsidizing bank loans to marketmakers improves social welfare.

It is assumed that, if a marketmaker borrows a quantity a, she can either "behave" or "shirk." If she behaves, she trades as described in Section 5 and makes profit (210). If she shirks, she does not provide liquidity (she passively match buyers and sellers) and steals a fraction $\phi \in (0, 1)$ of the capital initially borrowed. The opportunity to steal reflects, for instance, difficulties to monitor and precisely evaluate a marketmaker's trading strategy during financial disruptions.

Second Best

Suppose that marketmakers borrow an aggregate quantity $x \in [0, \bar{x}]$. If all marketmakers shirk, then social welfare is

$$W_e(0) + (1 - \phi)x + \phi x + \bar{x} - x = W_e(0) + \bar{x}. \tag{214}$$

The first term on the left-hand side reflects the fact that, when they shirk, marketmakers do not provide liquidity. The second term is the amount of capital

that banks can recover after the crash. The third term is the amount of capital stolen. The last term is the amount of capital which is not lent to marketmakers at time zero. If, on the other hand, all marketmakers behave, then social welfare is $W_e(x) + \bar{x} \geq W_e(0) + \bar{x}$. Hence, welfare is always improved when marketmakers behave. Moreover, it can be assumed without loss of generality that marketmakers behave when they are not allocated any capital ($x = 0$). This implies that, for welfare analysis, one can restrict attention to situations in which all marketmakers behave.

A feasible anonymous capital allocation is a loan size $x \in [0, \bar{x}]$ and a repayment $R \in \mathbb{R}_+$. An allocation is feasible if it satisfies the marketmaker's incentive compatibility and the bank's break-even constraint

$$\big(1 + F(x)\big)\, x - R\, x \geq \phi x \tag{215}$$

$$R\, x \geq x, \tag{216}$$

respectively (incentive compatibility implies limited liability (212)). In particular, any allocation $(x, R) \in [0, \bar{x}] \times \mathbb{R}_+$ such that $x = 0$ is feasible. The second-best program is to choose a feasible allocation maximizing $W_e(x)$, subject to (215) and (216) . One can show easily

Proposition 10. *If $\phi \in \big[F(\bar{x}), F(0)\big]$, then the second-best program is solved by $R = 1$ and $x = F^{-1}(\phi)$.*

In words, in this second-best allocation, a marketmaker makes the smallest repayment consistent with a bank breaking even, and aggregate marketmaking capital is the largest quantity consistent with a marketmaker behaving.

Implementation

The second best allocation is implemented by the market arrangement described earlier. The bank makes a take-it-or-leave-it offer to the marketmaker, consisting in a loan size $a \in [0, \bar{x}]$, and a repayment $R \in \mathbb{R}_+$, in order to maximize profit $(R-1)\, a$, subject to the marketmaker's incentive compatibility $\big(1+F(x)-R\big)a \geq \phi a$. Clearly, it is optimal for the bank to offer $R = 1 + F(x) - \phi$, which is the largest repayment consistent with the marketmaker behaving. After substituting this into the bank's objective, one can define an equilibrium as some $x^* \in [0, \bar{x}]$ such that

$$x^* \in \arg\max_{a \in [0, \bar{x}]} a\big(F(x^*) - \phi\big) + \bar{x}. \tag{217}$$

Clearly, if $\phi \in [F(\bar{x}), F(0)]$, the second best $x = F^{-1}(\phi)$ is an equilibrium.

Proposition 11. *If $\phi \in [F(\bar{x}), F(0)]$, there exists an equilibrium whose allocation is a second-best allocation.*

A simple way to improve the equilibrium of Proposition 11 is to subsidize bank loans. For instance, if a "government" levies lump-sum taxes in order to finance a small subsidy $\varepsilon > 0$ per unit lent, then the equilibrium aggregate marketmaking capital is $F^{-1}(\phi - \varepsilon) > F^{-1}(\phi)$, and welfare is improved.

The government ability to enforce tax payments is crucial for that result. Assume that agents choose to contribute to the subsidy. This improves aggregate welfare. Then, it would be optimal for an individual agent *not* to contribute and to free-ride on the welfare improvement.

F.3 Characterization of $F(\,\cdot\,)$ and $W_e(\,\cdot\,)$

Let us consider the constrained optimal allocations with inventory bound $M = m$ studied in Section A. Welfare in such an allocation can be written $W(m)$, for some increasing function $W(\,\cdot\,)$. This allocation is implemented in a competitive equilibrium in which time-zero aggregate marketmaking capital is set to

$$X(m) = \lambda_I(t_m^-)e^{-rt_m}m, \tag{218}$$

and where the gross return on marketmaking capital is

$$G(m) = \frac{\lambda_I(t_m^+)}{\lambda_I(t_m^-)}, \tag{219}$$

with

$$\begin{aligned}
\lambda_I(t_m^+) &= \frac{1-\delta_2}{r} + \frac{1}{r}\left(1 - \frac{\delta_1\gamma_d}{r+\rho+\gamma}\right)e^{-r\phi_2(m)} && (220)\\
\lambda_I(t_m^-) &= \lambda(t_m^+) - b(m) && (221)\\
t_m &= \psi(m), && (222)
\end{aligned}$$

where $\phi_i(\,\cdot\,)$, $i \in \{1,2\}$, and $\psi(\,\cdot\,)$ are defined in Lemma 2 and $b(\,\cdot\,)$ is defined in equation (114). For small $m > 0$, the derivative of $\lambda_m(t_m^-)$ with respect to m is

$$\begin{aligned}
&\left(\delta_2 - \frac{\delta_1\gamma_d}{r+\rho+\gamma}\right)\left(e^{-r\phi_2(m)} - e^{-(r+\rho)\phi_2(m)}\right)\phi_2'(m)\\
&\qquad + \left(\delta_2 - \frac{\delta_1(r+\rho+\gamma_d)}{r+\rho+\gamma}\right)e^{(r+\rho)\phi_1(m)}\phi_1'(m). \qquad (223)
\end{aligned}$$

Lemma 2 shows that $\phi_i(m) = \alpha\sqrt{m}(1+o(m))$ and that $\phi_i'(m) = \alpha(1+o(m))/(2\sqrt{m})$, for $i \in \{1,2\}$, and for some $\alpha > 0$. This, together with (223), implies $\lambda_m(t_m^-)$ is an increasing function for small m. Since $\psi(m)$ is decreasing, this implies that $X(m)$ is

a strictly increasing continuous function of m, for small m. Similarly, since $\lambda_I(t_m^+)$ is an strictly decreasing function of m, $G(m)$ is a strictly decreasing function for small m. Therefore, for small x, one can define

$$W_e(x) \equiv W\big(X^{-1}(x)\big) \tag{224}$$

$$F(x) \equiv G\big(X^{-1}(x)\big), \tag{225}$$

which concludes the proof.

References

Amihud, Yakov, and Haim Mendelson, Dealership Market: Marketmaking with Inventory, *Journal of Financial Economics*, 1980, *8*, 31–53.

Bernanke, Ben S., Clearing and Settlement during the Crash, *Review of Financial Studies*, 1990, *3*, 133–151.

Bernardo, Antonio E., and Ivo Welch, Liquidity and Financial Market Runs, *Quaterly Journal of Economics*, 2004.

Brady, Nicholas F., *The Presidential Task Force on Market Mechanisms*, Washington DC: US Government Printing Office, 1988.

Brémaud, Pierre, *Point Processes and Queues*, New-York: Springer-Verlag, 1981.

Demsetz, Harold, The Cost of Transacting, *The Quarterly Journal of Economics*, 1968, *82*, 33–53.

Diamond, Douglas W., and Philip H. Dybvig, Bank Runs, Deposit Insurance, and Liquidity, *Journal of Political Economy*, 1983, *91*, 401–417.

Duffie, Darrell, Nicolae Gârleanu, and Lasse H. Pedersen, "Marketmaking in Over-the-Counter Markets," Working Paper, Graduate School of Business 2001. Stanford University.

——, ——, and ——, "Valuation in Over-the-Counter Markets," Working Paper, Graduate School of Business 2001. Stanford University.

Fleming, Michael J., and Kenneth D. Garbade, When the Back-Office Moved to the Front Burner: Settlement Fails in the Treasury Market after 9/11, *Federal Reserve Bank of New-York Economic Policy Review*, 2002, *November*, 35–57.

Garman, Mark, Market Microstructure, *Journal of Financial Economics*, 1976, *3*, 257–275.

Goldstein, Michael A., and Kenneth A. Kavajecz, Trading Strategies during Circuit Breakers and Extreme Market Movements, *Journal of Financial Markets*, 2003, *Forthcoming.*

Greenberg, Maurice, Shake up the NYSE Specialist System or Drop it, *Financial Times*, October 10, 2003.

Greenwald, Bruce, and Jeremy Stein, The Task Force Report: The Reasoning Behind the Recommendations, *Journal of Economic Perspective*, 1988, *2*, 3–23.

——, and ——, Transactional Risk, Market Crashes, and the Role of Circuit Breakers, *Journal og Business*, 1991, *64*, 443–462.

Grossman, Sanford J., and Merton H. Miller, Liquidity and Market Structure, *Journal of Finance*, 1988, *43*, 617–637.

Hall, George, and John Rust, Middlemen versus Market Makers: A Theory of Competitive Exchange, *Journal of Political Economy*, 2003, *111*, 353–403.

Ho, Thomas, and Hans R. Stoll, Optimal Dealer Pricing under Trading Transactions and Return Uncertainty, *Journal of Financial Economics*, 1981, *9*, 47–73.

____ , and ____ , The Dynamics of Dealer Markets Under Competition, *Journal of Finance*, 1983, *38*, 1053–1074.

Hosios, Arthur J., On the Efficiency of Matching and Related Models of Search and Unemployment, *The Review of Economic Studies*, 1990, *57* (1), 279–298.

Kamien, Morton I., and Nancy L. Schwartz, *Dynamic Optimization: the Calculus of Variation and Optimal Control Theory in Economics and Management*, New-York: North-Holland, 1991.

Kreps, David M., and Jose A. Scheinkman, Quantity Precommitment and Bertrand Competition Yield Cournot Outcomes, *The Bell Journal of Economics*, 1983, *14*, 326–337.

Li, Yiting, Middlemen and private information, *Journal of Monetary Economics*, 1998, *42*, 131–159.

Longstaff, Francis, The Flight to Liquidity Premium in U.S. Treasury Bond Prices, *Journal of Business*, 2003. Forthcoming.

Marès, Arnaud, Market Liquidity and the Role of Public Policy, *Bureau of International Settlement Papers*, 2001, *12*, 385–401.

Masters, Adrian, " Efficiency of Intermediation in Search Equilibrium", SUNY Albany 2004. Working paper.

McAndrews, James J., and Simon M. Potter, Liquidity Effects of the Events of September 11, 2001, *FRBNY Economic Policy Review*, 2002, *November*, 59–79.

Mildenstein, Eckart, and Harold Schleef, The Optimal Pricing Policy of a Monopolistic Marketmaker in Equity Market, *Journal of Finance*, 1983, *38*, 218–231.

Moen, Espen R., Competitive Search Equilibrium, *Journal of Political Economy*, 1997, *105*, 385–411.

Mortensen, Dale T., and Randall Wright, Competitive Pricing and Efficiency in Search Equilibrium, *International Economic Review*, 2002, *43*, 1–20.

O'Hara, Maureen, and George S. Oldfield, The Microeconomics of Market Making, *Journal of Financial and Quantitative Analysis*, 1986, *21*, 361–376.

Parry, Robert T., The October '87 Crash Ten Years Later, *Federal Reserve Bank of San Francisco Economic Letter*, 1997, *97-32.*

Protter, Philip, *Stochastic Integration and Differential Equations*, New York: Springer-Verlag, 1990.

Rubinstein, Ariel, and Asher Wolinsky, Middlemen, *Quarterly Journal of Economics*, 1987, pp. 581–594.

Seierstad, Atle, and Knut Sydsæter, Sufficient Conditions in Optimal Control Theory, *International Economic Review*, 1977, *18*, 367–391.

Shevchenko, Andrei, Middlemen, *International Economic Review*, 2004, *45*, 1–25.

Shimer, Robert, "Contract in a Frictional Labor Market", Working Paper, Department of Economics 1995. MIT.

Shleifer, Andrei, and Robert W. Vishny, The Limits of Arbitrage, *The Journal of Finance*, 1997, *52*, 35.

Spulber, Daniel F., Marketmaking by Price-Setting Firms, *Review of Economic Studies*, 1996, *63*, 559–580.

Stoll, Hans R., The Supply of Dealer Services in Securities Markets, *Journal of Finance*, 1978, *33*, 1133–1151.

Sun, Yeneng, "On the Sample Measurability Problem in Modeling Individual Risks", Working Paper, Department of Mathematics and Center of Financial Engineering 2000. National University of Singapore.

Taylor, Angus E., and Robert W. Mann, *Advanced Calculus*, New-York: Wiley, John and Sons, 1983.

Vayanos, Dimitri, and Tan Wang, "A Theory of On-The-Run and Off-The-Run Bond Markets," Working Paper, MIT 2003.

Weill, Pierre-Olivier, "Liquidity Premia in Dynamic Bargaining Markets," Working Paper, Department of Economics 2002. Stanford University.

___ , "Essays on Liquidity in Financial Markets", Ph.D. Dissertation, Department of Economics 2004. Stanford University.

Wigmore, Barrie A., Revisiting the October 1987 Crash, *Financial Analysts Journal*, 1998, *54*, 36–48.