Macroeconomic Dynamics, **15** (Supplement 1), 2011, 119–144. Printed in the United States of America.
doi:10.1017/S1365100510000714

LIQUIDITY PROVISION IN CAPACITY-CONSTRAINED MARKETS

PIERRE-OLIVIER WEILL
University of California, Los Angeles
and
National Bureau of Economic Research

We study a competitive dynamic financial market subject to a transient selling pressure when market makers face a capacity constraint on their number of trades per unit of time with outside investors. We show that profit-maximizing market makers provide liquidity in order to manage their trading capacity constraint optimally over time: they use slack trading capacity early to accumulate assets when the selling pressure is strong in order to relax their trading capacity constraint and sell to buyers more quickly when the selling pressure subsides. When the trading capacity constraint binds, the bid–ask spread is strictly positive, widening and narrowing as market makers build up and unwind their inventories. Because the equilibrium asset allocation is constrained Pareto-optimal, the time variations in bid–ask spread are not a symptom of inefficient liquidity provision.

Keywords: Market Makers Inventory Management, Search, Time-Varying Bid–Ask Spread

1. INTRODUCTION

We study market makers' dynamic liquidity provision in a theoretical financial market subject to a transient selling pressure. We assume that market makers face a capacity constraint on their number of transactions with outside investors per unit of time. We show that, in equilibrium, profit maximizing market makers provide liquidity to outside investors in order to manage their trading capacity constraint optimally over time: they build up asset inventories early on when the selling pressure is strong and unwind their inventories later on when the pressure subsides. Although market makers are competitive and investors can contact them instantly, the bid–ask spread is strictly positive when the aggregate capacity constraint is biding. Consistent with empirical evidence, the spread widens and narrows as market makers accumulate and unwind their inventories. Furthermore, the equilibrium allocation is constrained Pareto-optimal, implying that, in our equilibrium, a widening of the spread is not evidence of inefficient liquidity provision.

I thank Philipp Kircher and the participants of the JMCB-FRBC conference on Liquidity in Frictional Markets. Address correspondence to: Pierre-Olivier Weill, 8283 Bunche Hall, Department of Economics, University of California, Los Angeles, Los Angeles, CA 90095, USA; e-mail: poweill@econ.ucla.edu.

© 2011 Cambridge University Press 1365-1005/11

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

Our model of a transient selling pressure follows Grossman and Miller (1988): we assume that at time zero, all asset holders receive a negative "liquidity shock," lowering their utility for the asset. Then the market slowly recovers as new high-valuation investors enter the economy progressively over time. In reality, such negative shock may occur at the level of one particular security, for instance when a stock is delisted from an index [see Greenwood (2005)], but it can also occur for all stocks at the same time, as during the market crash of October 87 or the mini crash of October 97. The subsequent recovery arises as new capital flows into the market.

As in the search models Weill (2007) and Lagos et al. (2008), we assume that all trades are intermediated by market makers who dislike holding assets in their inventories. Our main innovation relative to their search framework is our model of illiquidity: instead of assuming that investors take random times to contact market makers, we assume that investors can trade instantly at some quoted bid and ask prices. What makes the market illiquid, then, is the assumption that market makers face a capacity constraint on their number of trades per unit of time with outside investors.

A common critique of search-in-finance models is that, aside from some thin over-the-counter markets, in reality individual investors do not suffer long trading delays: they can always obtain a quote from, say, an online broker. Of course, if they don't like the quote, they may *choose* to delay their trade. This is exactly what goes on in our model, as investors can always buy and sell instantly at some equilibrium bid and ask prices. Although, from this perspective, our friction may be viewed as more realistic than the search friction, the equilibrium asset allocation turns out to be very similar to the one arising in search models. First the trading capacity constraint creates aggregate delays in reallocating assets from early sellers to late buyers, just as with a search friction. Second, as in search models, we show that market makers provide liquidity to outside investors by building up inventories early on, when the selling pressure is strong, and unwinding these inventories later on when the pressure subsides. Indeed, a market maker anticipates that, when the economy recovers from the crash and asset demand builds up, he will be able to sell assets to buyers in two ways. He can either buy assets from additional sellers, which uses up an additional unit of trading capacity; or he can sell directly from his own inventories, which does not use any additional trading capacity. Therefore, by accumulating inventories early on when the selling pressure is strong, a market maker effectively relaxes his trading capacity constraint when the economy recovers. On the aggregate, this speeds up the reallocation of assets from the early sellers to the late buyers, just as in search markets.

The main novel prediction of the present capacity-constrained model is that, even with competitive market makers, the bid–ask spread is positive when the aggregate trading capacity constraint is binding. By contrast, under the same competitive assumption, the search models of Duffie et al. (2005), Weill (2007), and Lagos et al. (2008) would all predict a zero bid–ask spread. Our positive spread arises precisely because, in contrast with the search framework, trading

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

delays are not exogenous to outside investors. Instead, the prices must adjust so that some investors find it optimal to delay their trades, and so that the aggregate order flow meets market makers' aggregate trading capacity constraint. Intuitively, the ask price needs to increase so that some buyers find it optimal to delay, and conversely the bid price needs to decrease so that some sellers delay as well. Taken together, this creates a wedge between the bid and the ask. Of course, from a market maker's perspective, this positive bid–ask spread does not constitute an arbitrage: indeed, market makers have no remaining trading capacity for additional buy-low–sell-high transactions with outside investors.

In our equilibrium, the bid–ask spread is time-varying: it widens and narrows down as market makers accumulate and unwind inventories. This prediction corroborates the evidence of Comerton-Forde et al. (2010) who show that, on the New York Stock Exchange (NYSE), stocks bid–ask spreads tend to go up and down with the inventories of their assigned specialists. As noted by these authors, such time variation in bid–ask spread are puzzling from the point of view of traditional inventory models of market making [see, among many others, Ho and Stoll (1981)], where market makers manage their inventories by changing the levels of their bid and ask, but always keep their spread the same.

In the broader search literature, our model is similar to the directed search model of Burdett et al. (2001): as in their model, although there is no "search problem," a capacity constraint distorts the equilibrium allocation. The main difference is that we consider a dynamic model where we assume away Burdett et al.'s coordination friction, and focus instead on the optimal dynamic management of the trading capacity constraint. Needless to say, it would be interesting to add a coordination friction in our dynamic model.

The remainder of the paper is organized as follows. In Section 2, we describe the economic environment, in Section 3 we solve for an equilibrium, and in Section 4 we study the comparative statics of the model.

2. THE ECONOMIC ENVIRONMENT

This section describes agents' preferences, the liquidity shock, and market makers' trading capacity constraint.

2.1. Market Makers and Investors

Time is continuous and runs forever. The economy is populated by a nonatomic continuum of infinitely lived and risk-neutral agents who discount the future at the same rate $r > 0$. An agent enjoys the consumption of a nonstorable numéraire good called "cash," with a marginal utility normalized to 1, and derives a linear utility flow θq from holding q units of an asset. There are three types of agents differing in their marginal utility θ for holding assets. The first two types are high- and low-valuation investors. The marginal utility θ of high-valuation investors is normalized to one, and the marginal utility of low-valuation in vestors is set to $1 - \delta$, for some $\delta \in (0, 1)$. The third type of agents are market makers. Market

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

makers represent financial intermediaries who trade actively in the market (NYSE specialists, market makers of NASDAQ and of other markets, banks, hedge funds), without being final holders of the assets. For this reason, we follow Weill (2007) and Lagos et al. (2010) in assuming that market makers do not derive any direct utility from holding assets; i.e., their θ is equal to zero.

There is one asset in positive supply $s \geq 0$. We assume that investors can hold either zero or one unit of the asset. Market makers, on the other hand, can hold any quantity of the asset but they face a short-selling constraint. We denote by

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

the asset inventory of a representative market maker at time $t \geq 0$.[1] As in Duffie et al. (2005), we can index investors by their preference type (high "h" or low "ℓ"), as well as their ownership status (owner "o" or nonowner "n"), so that the set of investors' types is $\mathcal{T} \equiv \{\ell o, hn, ho, \ell n\}$. In what follows we anticipate equilibrium behavior by calling an investor of type ℓo a "seller," and an investor of type hn a "buyer." The measure of type-σ investors is denoted

$$\mu_\sigma(t) \geq 0. \tag{2}$$

Because the asset has to be held either by investors or by market makers, we also have

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

2.2. Liquidity Shock

We assume that the market is hit by a negative transient liquidity shock, creating large and temporary selling pressure for the asset. A liquidity shock can be a marketwide event, affecting all stocks and investors at the same time, such as the market crashes of October 87 and October 97. Importantly, liquidity shocks can also occur at the level of individual stocks, for instance, when a stock is delisted from an index or an exchange [see Greenwood (2005)], forcing institutions, committed to hold the index to sell. Hendershott and Seasholes (2007) provide indirect evidence of stock-specific liquidity shock, showing that NYSE specialists occasionally take large long positions in their assigned stocks.

Our formalization of a transient liquidity shock is a natural continuous-time extension of the two-period liquidity shock model of Grossman and Miller (1988). We assume that, at time zero, the liquidity shock lowers investors' utility for holding the asset: for simplicity, we assume that all assets are held by low-valuation investors, and there are no high-valuation investors in the economy. This results in the initial distribution of types $\mu_{\ell o}(0) = s$ and $\mu_{hn}(0) = \mu_{ln}(0) = \mu_{ho}(0) = I(0)$. Then high-valuation investors enter the economy progressively over time. Namely, we assume that, at time $t \geq 0$, the measure $\mu_h(t)$ of high-valuation

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

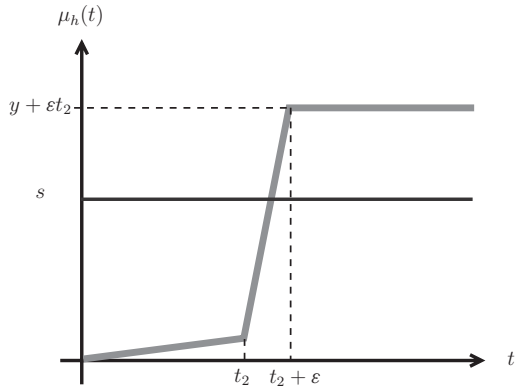


FIGURE 1. The time path of the measure $\mu_h(t)$ of high-valuation investors. It starts at zero, then increases at the low rate ε during $[0, t_2]$, and then at the high rate $1/\varepsilon$ during $[t_2, t_2+\varepsilon]$, before reaching its steady state level.

investors in the economy is equal to

$$\mu_h(t) = \begin{cases} \varepsilon t & t \in [0, t_2] \\ \varepsilon t_2 + y/\varepsilon(t - t_2) & t \in [t_2, t_2 + \varepsilon], \\ \varepsilon t_2 + y & t \in [t_2 + \varepsilon, \infty) \end{cases}$$

where $y > s$, $t_2 > 0$ is some exogenously given time, and $\varepsilon > 0$ is assumed to be small enough, in the sense that $\varepsilon < \min\{K, y/(2K), s/t_2\}$. The time path of $\mu_h(t)$ is illustrated in Figure 1. According to the figure, buyers enter the economy at a low rate early on when $t \in [0, t_2]$, and then at a very high rate when $t \in [t_2, t_2+\varepsilon]$. All buyers have entered the economy by time $t_2 + \varepsilon$.

Note that, as ε goes to zero, the recovery path $\mu_h(t)$ converges to a step function. Although we could work directly with this limiting case, we prefer to assume that $\varepsilon > 0$, so that our equilibrium has a well-defined bid–ask spread: we need a positive flow of high-valuation investors purchasing the asset to obtain a well-defined ask price, which will sometimes be different from the bid price at which low-valuation investors sell their asset.

In the context of this paper, the asset supply s is naturally interpreted as the size of the liquidity shock: it is the measure of assets that will be reallocated from low- to high-valuation investors. We end this section by noting that the distribution of types must satisfy the following accounting identities:

$$\mu_{\ell n}(t) + \mu_{\ell o}(t) = s, \tag{4}$$

$$\mu_{hn}(t) + \mu_{ho}(t) = \mu_h(t). \tag{5}$$

Equation (4) means that the total number of low-valuation investors must remain the same over time, whereas equation (5) means that the total number of high-valuation investors must be equal to $\mu_h(t)$.

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

2.3. Trading Technology

Our main innovation relative to the search-in-finance literature is our specification of the trading technology. Instead of making the "search" assumption that investors face exogenous delays in executing their trade, we assume (i) that investors can trade instantly, but (ii) that there is a marketwide capacity constraint on market makers' flow of trades with outside investors. The first assumption captures the intuitive notion that investors can instantly obtain a quote from a broker. The second assumption of a trading capacity constraint makes the market illiquid: in the face of a high enough selling pressure, market makers will not be able to simultaneously trade against all the orders that would arise if the capacity constraint were removed.[2] Thus, the trading capacity constraint creates delays in reallocating assets from low- to high-valuation investors, just as in the search literature. In contrast to the search literature, trade delays need to be part of investors' optimal trading strategies: thus, the bid and the ask will have to adjust, so some investors find it optimal to delay their trades.

Formally, let $u_\ell(t)$ and $u_h(t)$ be the aggregate flows of sell and buy orders originating from outside investors. By definition

$$u_\ell(t) \geq 0, \tag{6}$$

$$u_h(t) \geq 0. \tag{7}$$

The market aggregate trading capacity constraint is, then,

$$u_\ell(t) + u_h(t) \leq 2K \tag{8}$$

for some $K > 0$. Note that we do not impose any trading capacity constraint on the flow of trades *between market makers*.[3] The assumption means that professional traders have an easier time trading among each other than with outside investors. For instance, in addition to trading through the stock exchange, they can use the "upstairs" market for large transactions, as well as the services of interdealer brokers. Keep in mind, however, that limiting trades between market makers would not eliminate their incentives to hold inventories.[4]

In principle, one should describe how the total capacity $2K$ is distributed among market makers. However, because they can trade between each other without trading capacity constraint, we can aggregate market makers into one single representative competitive market maker, endowed with the aggregate capacity $2K$.

Our trading capacity constraint assumption is similar to the directed search assumption of Burdett et al. (2001), with one important difference. The similarity is that, in the directed search model, it does not take time for a buyer to find a seller, and a seller may not have the capacity to serve all buyers visiting him or her. The difference is that, in the directed search model, sellers are spatially separated: they cannot trade among each other after buyers make their one-time choice of which seller to visit. This creates a coordination friction, whereby in equilibrium some sellers do not have enough capacity to serve all their buyers, whereas other

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

sellers are left with some slack capacity. Our model, by contrast, assumes away coordination frictions by letting market makers trade among each other without trading capacity constraint: thus, in an equilibrium, either all capacity constraints will bind at the same time, or none of of them will bind. One may wonder whether the model remains interesting after the coordination friction is removed. Indeed, in the static environment of Burdett et al. (2001), removing the coordination friction would result a textbook model of Bertrand competition. The reason the present model remains interesting is because it is dynamic: market makers face a dynamic capacity-management problem that they solve using their inventory.

3. EQUILIBRIUM

We now construct an equilibrium under the following market arrangement. At each point in time, outside investors can buy the asset at the ask price, $A(t)$, and sell it at the bid price, $B(t)$. Market makers can trade with each others at some interdealer price, $p(t)$. Importantly, market makers' trade with outside investors are capped by the constraint (8). In the equilibrium we construct, market makers provide liquidity to outside investors: they buy assets early on from low-valuation investors and they resell these assets later on to high-valuation investors.

The construction is organized as follows. In Section 3.1, we make a guess on general properties of the equilibrium allocation, and of the corresponding trading strategies. In Section 3.2 we solve for equilibrium prices using some optimality conditions on investors and market makers. Finally, in Section 3.3, we verify that, given these prices, the conjectured trading strategies are indeed optimal.

3.1. Buffer Allocation

We first make the natural guess that, in an equilibrium, low-valuation investors will always find it optimal to sell the asset, and high-valuation to buy. Thus, the state variables $\mu_{\ell o}(t)$, $\mu_{hn}(t)$, $\mu_{ho}(t)$, and $\mu_{\ell n}(t)$ solve the system of ODEs

$$\dot{\mu_{\ell o}}(t) = -u_\ell(t), \tag{9}$$

$$\dot{\mu_{hn}}(t) = -u_h(t) + \dot{\mu}_h(t), \tag{10}$$

$$\dot{\mu_{ho}}(t) = u_h(t), \tag{11}$$

$$\dot{\mu_{\ell n}}(t) = u_\ell(t). \tag{12}$$

After differentiating (3), we find that the aggregate inventories of market makers solve the ODE

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

If market makers act as matchmakers and immediately resell to buyers the assets they purchase from sellers, then $u_\ell(t) = u_h(t)$ and $\dot{I}(t) = 0$. Otherwise, if $u_\ell(t) \neq u_h(t)$, then market makers are changing their inventory positions.

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

Next, we guess that market makers provide liquidity to outside investors: they build up inventories early on, when the selling pressure is high and the buying pressure is low, and unwind their inventories later on, when the buying pressure is high. Formally, we guess that the equilibrium asset allocation is characterized by breaking times $t_1 \leq t_2 \leq t_3 \leq t_4$, such that

$$u_\ell(t) = \varepsilon \quad \text{and} \quad u_h(t) = \varepsilon, \qquad t \in [0, t_1], \tag{13}$$

$$u_\ell(t) = 2K - \varepsilon \quad \text{and} \quad u_h(t) = \varepsilon, \qquad t \in [t_1, t_2], \tag{14}$$

$$u_\ell(t) = 0 \quad \text{and} \quad u_h(t) = 2K, \qquad t \in [t_2, t_3], \tag{15}$$

$$u_\ell(t) = K \quad \text{and} \quad u_h(t) = K, \qquad t \in [t_3, t_4], \tag{16}$$

$$u_\ell(t) = 0 \quad \text{and} \quad u_h(t) = 0, \qquad t \in [t_4, \infty). \tag{17}$$

During the first time interval, $[0, t_1]$, market makers act as pure matchmakers. Moreover, given our assumption that $K > \varepsilon$, the trading capacity constraint (8) allows them to match the entire flow ε of incoming buyers with sellers. Hence, during that interval, a buyer is able to purchase an asset immediately when he or she enters the economy, and $\mu_{hn}(t) = 0$.

During the second time interval, $[t_1, t_2]$, market makers continue to match the entire flow of incoming buyers with sellers (i.e., $\mu_{hn}(t) = 0$), but they use up all their slack capacity to build up inventories. That is, $\dot{I}(t) = u_\ell(t) - u_h(t) = 2K - 2\varepsilon > 0$, given our assumption that $K > \varepsilon$. Clearly, the inventory position of market makers at time t_2 is

$$I(t_2) = 2(K - \varepsilon)\Delta,$$

where $\Delta \equiv t_2 - t_1$. Note that, at time t_2, there is a measure εt_2 of high-valuation investors who all received an asset upon entering the economy, meaning that $\mu_{ho}(t_2) = \varepsilon t_2$. Therefore, the accounting identity (3) implies that the measure of sellers at time t_2 is

$$\mu_{\ell o}(t_2) = s - \mu_{ho}(t_2) - I(t_2) = s - \varepsilon t_2 - 2(K - \varepsilon)\Delta. \tag{18}$$

Starting at the beginning of the third time interval $[t_2, t_3]$, buyers enter the economy at the high rate y/ε. Given our assumption that $\varepsilon < y/(2K)$, the trading capacity constraint (8) makes it impossible for market makers to sell assets to all of the incoming buyers, so the trading capacity constraint continues to bind. In addition, because market makers' marginal utility for the asset is lower than that of sellers, efficiency intuitively requires that they unload all their inventories before matching more sellers with buyers. That is, $u_\ell(t) = 0$ and $u_h(t) = 2K$. Hence, all inventories are unloaded at time

$$t_3 = t_2 + \frac{I(t_2)}{2K} = t_2 + \left(1 - \frac{\varepsilon}{K}\right)\Delta. \tag{19}$$

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

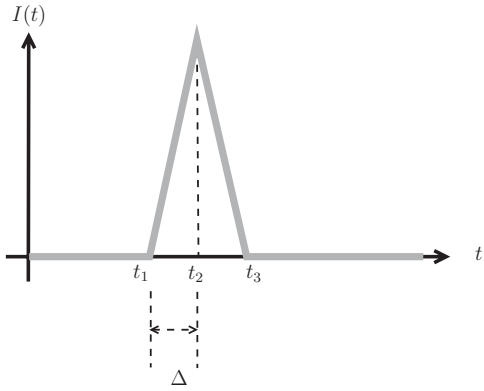


FIGURE 2. Inventories in a buffer allocation.

In the fourth time interval, $[t_3, t_4]$, market makers do not hold inventories anymore and match the remaining mass $\mu_{\ell o}(t_2)$ of sellers with buyers, at the maximum matching rate of K. This implies that all assets are transferred to high-valuation investors at time

$$\begin{aligned} t_4 &= t_3 + \frac{\mu_{\ell o}(t_2)}{K} \\ &= t_2 + \left(1 - \frac{\varepsilon}{K}\right)\Delta + \frac{s - \varepsilon t_2 - 2(K-\varepsilon)\Delta}{K} \\ &= t_2 + \frac{s - \varepsilon t_2}{K} - \left(1 - \frac{\varepsilon}{K}\right)\Delta, \end{aligned} \tag{20}$$

where the second equality follows by plugging equations (18) and (19) into the first equation. After time t_4, all assets are in the hands of their final holders, in the sense that $\mu_{\ell o}(t) = I(t) = 0$, and $\mu_{ho}(t) = s$.

Equations (13) and (14) provide a one-parameter family of asset allocation indexed by Δ, the length of the interval during which market makers build up their inventories. In what follows, we will call such an allocation a "buffer allocation." Figure 2 shows the time path of market makers' inventories in such an allocation.

Note that Δ is bounded above. Clearly, Δ must be smaller than t_2, but the upper bound may be tighter: for instance if K is very large, then if market makers start accumulating assets too early, their inventories at time t_2 will be greater than the asset supply, which would violate feasibility. The upper bound of Δ is easily found by requiring that $\mu_{\ell o}(t_2)$ be positive. Using equation (18), this condition yields

$$\Delta \in [0, \bar{\Delta}], \text{ where } \bar{\Delta} = \min\left\{t_2, \frac{s - \varepsilon t_2}{2(K-\varepsilon)}\right\}. \tag{21}$$

In Appendix A.1, we verify that

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

PROPOSITION 1 (Feasibility). *For all* $\Delta \in [0, \bar{\Delta}]$*, equations (13)–(17) and ODEs (9)–(12) define a feasible asset allocation, in the sense that the initial conditions and the constraints (1) through (8) are all satisfied.*

Equation (20) reveals that, in a buffer allocation, market makers speed up the allocation of assets to high-valuation investors. That is, starting to accumulate inventories at time $t_1 = t_2 - \Delta$, reduces the time to allocate assets by $(1 - \varepsilon/K)\Delta$ units of time. This is because inventory accumulation at the beginning allows market makers to relax their trading capacity constraints when the flow of buyers is large. Indeed, whereas they can match buyers with sellers at rate K, they can match buyers with their inventories at the higher rate of $2K$.

3.2. A Heuristic Derivation of Prices

We proceed with a heuristic characterization of the price paths $B(t)$, $p(t)$, and $A(t)$ that must prevail for the equilibrium allocation to be a buffer allocation. We derive a series of restrictions on prices using agents' optimality condition. We use these restrictions to guess candidate equilibrium prices.

Restrictions implied by market makers' optimality. First, for market makers' trading strategy to be optimal, we must have that

$$A(t) = B(t) = p(t) \quad \text{during } [0, t_1] \text{ and } [t_4, \infty), \tag{22}$$

because market makers' trading capacity constraint is slack. Indeed, if $p(t) < A(t)$, then market makers could use slack capacity to buy assets from other market makers, sell to outside investors, and make more profit. Similarly, if $p(t) > B(t)$, then a market maker could use slack capacity to buy from outside investors and sell to outside investors. Next, we argue that

$$p(t) = \frac{A(t) + B(t)}{2} \quad \text{during } [t_1, t_2] \text{ and } [t_3, t_4]. \tag{23}$$

That is, when the trading capacity constraint is binding and market makers trade with outside investors, then the interdealer price is equal to the midpoint between the bid and the ask price.

To see why, recall that the representative market maker's prescribed trading strategy is to buy $u_\ell(t)$ at $B(t)$ and sell $u_h(t)$ at $A(t)$, where $u_\ell(t) + u_h(t) = 2K$. Now consider the following deviation during a small time interval $[t, t + dt]$: buy $2K$ at $B(t)$, sell $u_h(t) - u_\ell(t) + 2K$ at $p(t)$, and continue with the prescribed trading strategy thereafter. Both trades use all capacity and leave the market maker's inventory position unchanged. Because the prescribed strategy must be weakly better than the deviation, we have

$$\begin{aligned} &A(t)u_h(t) - B(t)u_\ell(t) \geq p(t)\,(u_h(t) - u_\ell(t) + 2K) - B(t)2K \\ &\Leftrightarrow 2u_h(t)p(t) \geq u_h(t)\,(A(t) + B(t)) \Leftrightarrow p(t) \geq \frac{A(t) + B(t)}{2}, \end{aligned}$$

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

where the second line follows by substituting $u_h(t) + u_\ell(t) = 2K$ into the first line and rearranging.[5]

The reverse inequality obtains by considering the deviation of buying $2K + u_\ell(t) - u_h(t)$ at $p(t)$, and selling $2K$ at $A(t)$. Finally, we note that

$$rp(t) = \dot{p}(t) \quad \text{during } [t_1, t_3]. \tag{24}$$

This condition, which is standard from the literature on speculative trading, ensures that market makers are willing to hold a positive-inventory position. It says that the price must grow at a rate that compensates market makers for the time value of the cash spent on inventories. If $\dot{p}(t)/p(t) > r$, then market makers would demand an infinite amount of inventory in the interdealer market, which cannot be the basis of an equilibrium. If $\dot{p}(t)/p(t) < 0$, on the other hand, market makers would prefer a zero-inventory position.

Restrictions using high-valuation optimality. During $[t_2, t_4]$, some high-valuation investors buy the asset immediately, and others choose to delay their purchase until time t_4. Thus, in order for delays to be individually optimal, the following indifference condition must hold:

$$-A(t) + \frac{1}{r} = \frac{1}{1 + rdt}\left(-A(t + dt) + \frac{1}{r}\right),$$

which says that the net utility of buying at t and holding the asset forever, on the left-hand side, must be equal to the net utility of buying at $t + dt$. Rearranging and neglecting all second-order terms, one obtains the familiar asset pricing equation $rA(t) = 1 + \dot{A}(t)$.

Now consider what happens during $[t_4, \infty)$. For a high-valuation nonowner, the net utility of buying at t, holding for dt, and reselling at $t + dt$ must be negative:

$$-A(t) + \frac{1}{1 + rdt}\left(dt + B(t + dt)\right) \leq 0 \Leftrightarrow rA(t) \geq 1 + \dot{A}(t),$$

given that $A(t) = B(t)$ and neglecting second-order terms. Similarly, for a high-valuation owner, the net utility of selling at t and buying back at $t + dt$ must be negative, leading to the opposite inequality. Taken together, we obtain again that $rA(t) = 1 + \dot{A}(t)$. Imposing the transversality condition $\lim_{t\to\infty} e^{-rt}A(t) \to 0$,[6] we obtain that

$$A(t) = \frac{1}{r} \quad \text{during } [t_2, \infty). \tag{25}$$

Restrictions implied by low-valuation optimality. In a buffer allocation, a low-valuation investor must find it optimal to hold the asset during $[0, t_4]$ and sell the asset at the bid price during $[0, t_2] \cup [t_3, \infty]$. Thus, during $[0, t_2] \cup [t_3, t_4]$, a low-valuation investor must be indifferent between selling at time t and holding the

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

asset for a small time interval of length dt and selling at time $t + dt$. Proceeding as in the preceding section, we obtain that

$$r B(t) = (1 - \delta) + \dot{B}(t) \quad \text{during } [0, t_2] \text{ and } [t_3, t_4]. \tag{26}$$

At any time $t \in [t_2, t_3)$, a low-valuation investor must prefer to delay its sale until t_3:

$$B(t) \leq \int_t^{t_3} (1 - \delta) e^{-r(z-t)}\, dz + e^{-r(t_3-t)} B(t_3), \tag{27}$$

with equality at time $t = t_2$.

The guess. We now formulate a guess that is consistent with restrictions (22) through (27). First, imposing the indifference condition (26) during $[0, t_4]$ and using $B(t) = A(t) = 1/r$ for $t \geq t_4$, we obtain

$$B(t) = \int_t^{t_4} (1 - \delta)\, e^{-r(z-t)}\, dz + \frac{e^{-r(t_4-t)}}{r} \quad \text{during } [0, t_4], \tag{28}$$

$$B(t) = \frac{1}{r} \quad \text{during } [t_4, \infty). \tag{29}$$

Next, we guess that the interdealer price is

$$p(t) = B(t) \quad \text{during } t \in [0, t_1], \tag{30}$$

$$p(t) = e^{-r(t_3-t)} p(t_3) \quad \text{during } t \in [t_1, t_3], \tag{31}$$

$$p(t) = \int_t^{t_4} e^{-r(z-t)} \left(1 - \frac{\delta}{2}\right) dz + \frac{e^{-r(t_4-t)}}{r} \quad \text{during } t \in [t_3, t_4], \tag{32}$$

$$p(t) = \frac{1}{r} \quad \text{during } [t_4, \infty). \tag{33}$$

Equations (30) and (33) follow from (22), and equation (31) from (24). Equation (32) obtains by plugging into (23) the formulas (25) and (28) for $A(t)$ and $B(t)$. Last, we guess that the ask price is

$$A(t) = B(t) \quad \text{during } t \in [0, t_1], \tag{34}$$

$$A(t) = 2p(t) - B(t) \quad \text{during } t \in [t_1, t_2), \tag{35}$$

$$A(t) = \frac{1}{r} \quad \text{during } t \in [t_2, \infty). \tag{36}$$

Equation (35) follows from $p(t) = 1/2(A(t) + B(t))$. Note that the above construction implies that $p(t)$ is continuous during $[0, t_1)$ and $(t_1, \infty]$, but does not guarantee continuity at $t = t_1$. We now argue that, if at equilibrium $t_1 > 0$, then $p(t)$ must be continuous at $t = t_1$. Indeed, if $p(t)$ jumps up, then a market maker could make more profit by buying the asset on the interdealer market

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

before the jump, and reselling after the jump. If $p(t)$ jumps down, then because $A(t) = 2p(t) - B(t)$ and $B(t)$ is continuous, it follows that $A(t)$ jumps down as well. In that case, when t is close to t_1, a high-valuation investor would find it optimal to delay his purchase in order to buy at a lower price, meaning that the conjectured trading strategy is not optimal.

Combining equation (30) with equation (31), one finds that $p(t)$ is continuous at $t = t_1$ if and only if $J(\Delta) = 0$, where

$$J(\Delta) \equiv \left(1 - \frac{\delta}{2}\right) e^{-2r(1-\varepsilon/K)\Delta} - \frac{\delta}{2} e^{-r(s-\varepsilon t_2)/K} - (1-\delta) e^{r\varepsilon/K\Delta}. \tag{37}$$

3.3. Verification

We now proceed to verify that the above conjecture is the basis of an equilibrium. We show

PROPOSITION 2 (Equilibrium). *There exists an equilibrium whose allocation is a buffer allocation. The associated parameter Δ^* is the unique solution of $J(\Delta^*) = 0$, or $\Delta^* = t_2$ and $J(t_2) > 0$. The associated prices are given by (28) through (36).*

The equilibrium prices are illustrated in Figure 3. Maybe the most important feature of our equilibrium is that, even with perfectly competitive market makers, the bid–ask spread $A(t) - B(t)$ is strictly positive during $[t_1, t_4]$, when the trading capacity constraint is binding. Although buyers can contact market makers instantly, they cannot bid the ask price $A(t)$ down to the interdealer price $p(t)$, and vice versa for sellers and the bid price $B(t)$. This is because outside investors have to compensate market makers for the shadow cost of using a unit of capacity when transacting. The flip side of the argument is that the positive bid–ask spread does not constitute arbitrage: indeed the binding trading capacity constraint prevents market makers from making unbounded profit by buying additional shares at the bid price $B(t)$ and selling them at the ask price $A(t)$.

Note that, according to our capacity-constrained model, a positive bid–ask spread obtains when there is a large demand for transacting on one side of the market, and the transaction demand on the other side is also large, as in $[t_2, t_3]$, or expected to be large soon, as in $[t_1, t_2]$. The bid–ask spread is zero at the beginning, when the market is "one-sided": the selling pressure is high but there is not much demand to transact from the other side of the market.

An exotic feature of the equilibrium is that the ask price $A(t)$ jumps up at time t_2, when the flow of incoming buyers becomes large relative to the trading capacity constraint. Note that, at equilibrium, a market maker finds it optimal to sell assets until the last instant before the jump. One might wonder why a market maker does

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

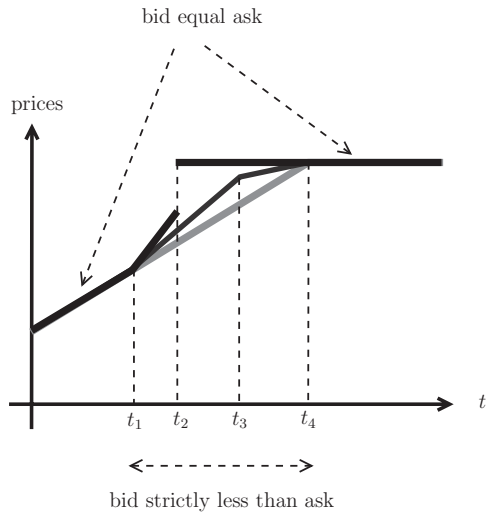


FIGURE 3. The bid (bottom), the interdealer price (middle), and the ask (top).

not instead wait for the jump in order to sell the asset at a much higher ask price. In fact, a market maker cannot sell more assets after the jump, because he or she already uses all his or her trading capacity. Instead, he or she would have to wait until time t_4, when his or her trading capacity constraint is slack.

4. IMPLICATIONS

4.1. Efficiency

In this section we show that the equilibrium allocation is Pareto optimal, subject to the aggregate trading capacity constraint. We first define a *feasible allocation* as a collection of $\mu_\sigma(t)$, $u_h(t)$, and $u_\ell(t)$ satisfying the initial conditions, the ODEs (9) through (12), and the positivity restrictions (1), (2), (6), and (7), as well as the trading capacity constraint (8).[7] To solve for constrained Pareto optimal allocations, recall that agents derive utility from two goods: the dividend flow of the asset, and the numéraire good. Because the utility for the numéraire good is linear and the same for all agents, it follows that, in all Pareto optimal allocations, the asset allocation maximizes the utilitarian criterion, the equally weighted sum of agents' utility:

$$\int_0^\infty (\mu_{ho}(t) + (1-\delta)\mu_{\ell o}(t))\, e^{-rt}\, dt, \tag{38}$$

among all feasible asset allocations [see, e.g., Chap. 10 of Mas-Colell et al. (1995)]. In Appendix A.3, we set up and solve the planner's problem using standard Hamiltonian techniques, and we show that:

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

PROPOSITION 3 (Efficiency). *The equilibrium allocation of Proposition 2 is Pareto optimal.*

There is a natural correspondence between the co-state variables of the Planner's Hamiltoninan and the equilibrium prices of Proposition 2: for instance, the interdealer price is equal to the co-state variable on the ODE for $I(t)$, which is the marginal social value of increasing market makers' inventories.

Proposition 3 is similar to Weill (2007) and Lagos et al. (2010), who also show that competition among market makers results in efficient allocation of assets. The main difference is that, given their specification of the search technology, efficiency requires a zero bid–ask spread. In our capacity-constrained framework, by contrast, the bid–ask spread can be positive. Thus, contrary to a commonly held view [see, e.g., the empirical study of Christie and Schultz (1998), Christie and Schultz (1998)], a positive bid–ask spread is not a symptom of inefficient liquidity provision.

4.2. Comparative Statics

In what follows we provide comparative static exercises. To simplify the algebra, and given that we are interested in results that hold for ε small enough, we focus on the limiting case $\varepsilon = 0$.

The amount of liquidity provided. A natural measure of the amount of liquidity provided is the length Δ of the inventory accumulation period. With Proposition 2, Δ can be written as $F(K, r, \delta, s)$, for some continuous function $F(\cdot)$. The following proposition provides some natural comparative statics.

PROPOSITION 4 (Comparative Statics). *Suppose* $\varepsilon = 0$ *and let* $x = (K, r, \delta, s)$ *be a vector of exogenous parameters. If* $F(x) < t_2$, *then* $F(\cdot)$ *is differentiable at x, with partial derivatives*

$$\frac{\partial F}{\partial K} < 0, \quad \frac{\partial F}{\partial r} < 0, \quad \frac{\partial F}{\partial \delta} > 0, \quad \frac{\partial F}{\partial s} > 0.$$

If, on the other hand, $F(x) = t_2$ and $J(F(x)) > 0$, then market makers start accumulating inventory at time zero and, locally, $F(\cdot)$ does not depend on the exogenous parameters x. Proposition 4 shows that the inventory-accumulation period is shorter if the capacity K is relaxed. Because a larger discount rate r and a smaller holding cost δ both reduce the present value of future asset transfers, either results in a shorter inventory-accumulation period. Last, the inventory-accumulation period is larger when s is larger: intuitively, a larger s means that the liquidity shock is larger, in the sense that the market needs to reallocate more assets from low- to high-valuation investors.

Another measure of the amount of liquidity provided is the maximum inventory position $I(t_2) = 2K\Delta^* = 2KF(K, r, \delta, s)$ held by market makers. Clearly, the sign of the partial derivatives of $I(t_2)$ with respect to (r, δ, s) is the same as that

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

of $F(K, r, \delta, s)$. The partial derivative with respect to K, on the other hand, is ambiguous: indeed, when K increases, market makers accumulate inventories for a shorter time but at a higher rate. The next proposition settles the issue:

PROPOSITION 5 (The Maximum Inventory Position). *Suppose $\varepsilon = 0$ and let $x = (K, r, \delta, s)$ be a vector of exogenous parameters. If $F(x) < t_2$, then the partial derivative of $2KF(x)$ with respect to K is positive.*

One may wonder what happens to $I(t_2)$ in the Walrasian limit, when K goes to infinity. There are two effects going in opposite directions: as K goes to infinity Δ goes to zero and, at the same time, K goes to infinity. In the $\varepsilon = 0$ case, we know from Proposition 5 that the maximum inventory position increases in K, so it has a nonzero limit as K goes to infinity. This result, however, is special to the $\varepsilon = 0$ case. If $\varepsilon > 0$, then the trading capacity constraint ceases to bind if $K > y/\varepsilon$: even around t_2, it is always possible to match all of the incoming flow y/ε of high-valuation investors with some low-valuation investors. Thus, for large but finite values of K, market makers do not accumulate any inventories. Therefore, when $\varepsilon > 0$ and as K goes to infinity, the maximum-inventory position clearly converges to zero.[8]

Illiquidity measures. In this section we study how traditional measures of illiquidity vary over time and depend on the market trading capacity K.

We start with a study of the bid–ask spread, $\sigma(t) \equiv A(t) - B(t)$. Clearly, the spread is zero when $t \leq t_1$ and $t \geq t_4$, and positive in between. When $t \in [t_1, t_2)$, $A(t) = 2p(t) - B(t)$, so the spread is $\sigma(t) = 2(p(t) - B(t))$. Combining with the ODEs $rB(t) = (1 - \delta) + \dot{B}(t)$ and $rp(t) = \dot{p}(t)$, we obtain

$$r\sigma(t) = -2(1-\delta) + \dot{\sigma}(t) \Rightarrow \sigma(t) = \frac{2(1-\delta)}{r}\left(e^{r(t-t_1)} - 1\right) \text{ during } [t_1, t_2),$$

after integrating the ODE and using the initial condition $\sigma(t_1) = 0$. In the second phase, when $t \in [t_2, t_4]$, we have $A(t) = 1/r$ and $rB(t) = 1 - \delta + \dot{B}(t)$, so $\sigma(t) = A(t) - B(t)$ satisfies

$$r\sigma(t) = \delta + \dot{\sigma}(t) \Rightarrow \sigma(t) = \frac{\delta}{r}\left(1 - e^{-r(t_4-t)}\right) \text{ during } [t_2, t_3],$$

after integrating the ODE and using the terminal condition $\sigma(t_4) = 0$. The resulting time path of the spread is illustrated in Figure 4. The spread is initially zero, then widens when market makers build up inventories during $[t_1, t_2]$, narrows as they unwind their inventories during $[t_2, t_3]$, and converges to zero during $[t_3, t_4]$. Consistent with this basic prediction of our capacity-constrained model, Comerton-Forde et al. (2010) show that, on the NYSE, specialists quote wider bid–ask spreads when they have larger inventories. The authors argue that the finding of a negative relationship between illiquidity, measured by the bid–ask

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

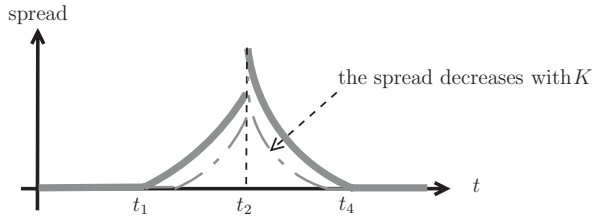


FIGURE 4. The bid–ask spread.

spread, and inventory levels is broadly consistent with recent models of market-making based on capital constraints [Gromb and Vayanos (2002), Brunnermeier and Pedersen (2009)]. One difficulty with this interpretation, however, is that these models do not predict any variation in the bid-ask spread. Instead, they all predict that the bid–ask spread is zero. The present paper provides an alternative explanation, based on a trading capacity constraint: in our equilibrium, the bid–ask spread is positive, and it goes up and down together with market makers' inventory levels.

Next, we study the impact of increasing capacity, K, on the size of the bid–ask spread:

PROPOSITION 6 (Capacity and Spread). *A marginal increase in capacity strictly decreases the bid–ask spread whenever it is positive, and keeps it equal to zero otherwise.*

This proposition suggest that improvements in trading technology should be associated with a reduction in the bid–ask spread: this prediction is broadly consistent with the secular decline in transaction cost documented in Jones (2002).

Next, we turn to the price discount conceded by a seller, the deviation $1/r - B(t)$ of the bid price from its steady state value. This natural measure of illiquidity is proposed, for instance, by Brunnermeier and Pedersen (2009). The discount is large at the beginning and decreases over time as new high-valuation investors flow into the market.

PROPOSITION 7 (Capacity and Price Discount). *A marginal increase in capacity decreases the price discount,* $1/r - B(t)$*, at each time* $t < t_4$.

Last, we study the impact of capacity K on what Black (1971) called "price resiliency" – the speed with which an asset price recovers from a random shock. In the present paper, a natural measure of resiliency is the time t_4 at which the price reaches its steady-state value of $1/r$.

PROPOSITION 8 (Capacity and Resiliency). *A marginal increase in capacity decreases* t_4.

We thus find the intuitive result that a greater capacity makes the market more resilient, as assets are reallocated faster to their final holders.

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

5. CONCLUSION

This paper studies the liquidity provision of market makers who face a capacity constraint on their number of trades with outside investors per unit of time. Similarly to search models, the capacity constraint creates delays in reallocating assets among investors, and these delays are mitigated by market makers' optimal inventory strategies. The key difference with search models is that investors can trade instantly, implying that the price must adjust so that investors find it optimal to delay their trade when the aggregate trading capacity constraint is binding. We show that, consistent with empirical evidence, this creates a time-varying bid–ask spread, widening and narrowing as market makers build up and unwind their inventories. We show that the positive bid–ask spread is consistent with an efficient asset allocation, and that it is narrower in markets with larger capacity.

NOTES

1. Instead of the above short-selling constraint, we could assume that a market maker incurs a short-selling cost $c > 1$ per unit of negative inventory. This could represent the dividend flow that a market maker would need to pay if he sold an asset he did not own. More generally, risk aversion and funding constraints make short positions more costly, than long positions. Consistent with this view, Hendershott and Seasholes (2007) find that NYSE specialists short positions tend to be smaller than their long positions.

2. For instance, after the "mini-crash' of October 1997, the Securities and Exchanges Commission (SEC) reported that "brokerdealers' web servers had reached their maximum capacity to handle simultaneous users" (SEC Staff Legal Bulletin No.8 at http:www.sec.govinterps/legalslbmr8.htm).

3. Duffie et al. (2005) make a similar assumption of frictionless trade among dealers.

4. In fact, if market makers start with zero inventories and have heterogenous endowment of capacities, then in equilibrium they do not trade between each other. Indeed, by homogeneity, an optimal trading strategy of an individual market maker is proportional to the aggregate trade of the market-making sector. In particular, an individual market maker's net trade with other market makers is zero.

5. It is important to note that our argument does not rely on there being a representative market maker. Indeed, suppose that market makers have heterogenous capacity. The trading strategies associated with the candidate equilibrium allocation generate an aggregate profit flow of $A(t)u_h(t) - B(t)u_\ell(t)$. Now consider changing each market makers strategy so that, on the aggregate, market makers purchase $2K$ at $B(t)$, sell $u_h(t) - u_\ell(t) + 2K$ at $p(t)$, and keep their "end of period" inventory the same. The associated aggregate profit flow must be less than the one generated by the equilibrium allocation, leading to the same conclusion that $p(t) \geq (A(t) + B(t))/2$.

6. Lagos et al. (2010) prove the necessity of the transversality condition in a closely related economic environment.

7. Our definition rules out transfers of asset from high-valuation to low-valuation investors: for completeness, at the end of the proof of Proposition 3, we verify that such transfers are indeed inefficient.

8. Note that this result cannot be derived using the calculations shown in the paper. Indeed, because we assume $K < y/(2\varepsilon)$, the present calculations do not address the limiting behavior of the economy as $K \to y/\varepsilon$. One can show that, as $K \to y/\varepsilon$, the maximum inventory position converges indeed to zero. These calculations are available from the author upon request.

9. The jump condition of $\lambda_h(t)$ is the opposite of that of $\lambda_\ell(t)$ and $\lambda_I(t)$ because $\lambda_h(t)$ enters the Hamiltonian with a minus sign.

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

REFERENCES

Black, Fischer (1971) Towards a fully automated exchange, part i. *Financial Analyst Journal* 27, 29–34.

Brunnermeier, Markus K. and Lasse H. Pedersen (2009) Funding liquidity and market liquidity. *Review of Financial Studies* 22, 2201–2238.

Burdett, Kenneth, Shouyong Shi, and Randall Wright (2001) Pricing and matching with frictions. *Journal of Political Economy* 109, 1060–1085.

Christie, William G. and Paul H. Schultz (1998) Dealer markets under stress: The performance of nasdaq market makers during the November 15, 1991, market break. *Journal of Financial Services Research* 13, 205–229.

Comerton-Forde, Carole, Terry Hendershott, Charles M. Jones, Pamela Moulton, and Mark S. Seasholes (2010) Time variation in liquidity: The role of market maker inventories and revenues. *Journal of Finance* 65, 295–331.

Duffie, Darrell, Nicolae Gârleanu, and Lasse H. Pedersen (2005) Over-the-counter markets. *Econometrica* 73, 1815–1847.

Greenwood, Robin (2005) Short and long-term demand curves for stocks: Theory and evidence on the dynamics of arbitrage. *Journal of Financial Economics* 75, 607–650.

Gromb, Denis and Dimitri Vayanos (2002) Equilibrium and welfare in markets with financially constraint arbitrageurs. *Journal of Financial Economics* 66, 361–407.

Grossman, Sanford J. and Merton H. Miller (1988) Liquidity and market structure. *Journal of Finance* 43, 617–637.

Hendershott, Terry and Mark S. Seasholes (2007) Marketmaker inventories and stock prices. *American Economic Review Papers and Proceedings* 97, 210–214.

Ho, Thomas and Hans R. Stoll (1981) Optimal dealer pricing under trading transactions and return uncertainty. *Journal of Financial Economics* 9, 47–73.

Jones, Charles M. (2002) A Century of Stock Market Liquidity and Trading Costs. Working paper, Columbia Graduate School Business.

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill (2008) Crashes and Recovered in Liquidity Markets. Working paper, NewYork University, University of California, Irvine, University of California, Los Angeles.

Lagos, Ricardo, Guillaume Rocheteau, and Pierre-Olivier Weill (2010) Crises and Liquidity in OTC Markets. Working paper, New York University, University of California, Irvine, University of California, Los Angeles.

Mas-Colell, Andreu, Michael D. Whinston, and Jerry R. Green (1999) *Microeconomic Theory*. Oxford, UK: Oxford University Press.

Seierstad, Atle and Knut Sydsæter (1977) Sufficient conditions in optimal control theory. *International Economic Review* 18, 367–391.

Weill, Pierre-Olivier (2007) Leaning against the wind. *Review of Economic Studies* 74, 1329–1354.

APPENDIX: PROOFS

A.1. PROOF OF PROPOSITION 1

First, the initial conditions are clearly satisfied. Second, by construction, the aggregate buy and sell order flows are $u_h(t)$ and $u_\ell(t)$, and they satisfy constraints (7), (6), and (8). Constraint (3) is verified at $t = 0$ because of the initial condition, as well as at all subsequent times $t > 0$ because, after adding up ODEs (9), (11), and (13), we have that $\dot{\mu}_{\ell o}(t) + \dot{\mu}_{ho}(t) + \dot{I}(t) = 0$. Similar reasoning implies that constraints (5) and (4) are also

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

satisfied. The short-selling constraint (1) is also satisfied by construction. The last thing to verify is that $\mu_\sigma(t) \geq 0$ for all $\sigma \in \{\ell o, \ell n, ho, hn\}$. Because $u_\ell(t) \geq 0$ and $u_h(t) \geq 0$, it follows that $\mu_{\ell n}(t)$ and $\mu_{ho}(t) \geq 0$. Now consider $\mu_{\ell o}(t)$. Direct calculations show that

$$\begin{aligned}
&\mu_{\ell o}(t) = s - \varepsilon t, && t \in [0, t_2],\\
&\mu_{\ell o}(t) = \mu_{\ell o}(t_2), && t \in [t_2, t_3],\\
&\mu_{\ell o}(t) = \mu_{\ell o}(t_2) - K(t - t_3), && t \in [t_3, t_4],\\
&\mu_{\ell o}(t) = 0. && t \geq t_4.
\end{aligned}$$

Our assumption that $\varepsilon t_2 < s$ implies that $\mu_{\ell o}(t) \geq 0$ at all times. Finally, consider $\mu_{hn}(t)$. Direct calculations show that, if $t_3 < t_2 + \varepsilon$,

$$\begin{aligned}
&\mu_{hn}(t) = 0, && t \in [0, t_2],\\
&\mu_{hn}(t) = \tfrac{y}{\varepsilon}(t - t_2) - 2K(t - t_2), && t \in [t_2, t_3],\\
&\mu_{hn}(t) = \tfrac{y}{\varepsilon}(t - t_2) - 2K(t_3 - t_2) - K(t - t_3), && t \in [t_3, t_2 + \varepsilon],\\
&\mu_{hn}(t) = y - 2K(t_3 - t_2) - K(t - t_3) && t \in [t_2 + \varepsilon, t_4],\\
&\mu_{hn}(t) = y - s + \varepsilon t_2 && t \geq t_4.
\end{aligned}$$

On the other hand, if $t_3 \geq t_2 + \varepsilon$,

$$\begin{aligned}
&\mu_{hn}(t) = 0, && t \in [0, t_2],\\
&\mu_{hn}(t) = \tfrac{y}{\varepsilon}(t - t_2) - 2K(t - t_2), && t \in [t_2, t_2 + \varepsilon],\\
&\mu_{hn}(t) = y - 2K(t - t_2), && t \in [t_2 + \varepsilon, t_3],\\
&\mu_{hn}(t) = y - 2K(t_3 - t_2) - K(t - t_3) && t \in [t_3, t_4],\\
&\mu_{hn}(t) = y - s + \varepsilon t_2 && t \geq t_4.
\end{aligned}$$

Our assumptions that $\varepsilon < y/2K$ and $y > s$ imply that $\mu_{hn}(t) \geq 0$.

A.2. PROOF OF PROPOSITION 2

Because a buffer allocation is feasible by construction, the only thing that has to be verified is that agents' trading strategies are indeed optimal.

We start with the representative market maker's problem, which is to maximize

$$\int_0^\infty [A(t)u_h(t) - B(t)u_\ell(t) - p(t)u(t)]\, e^{-rt}\, dt,$$

with respect to some time path $[u_h(t), u_\ell(t), u(t), I(t)]$, and subject to the ODE (13), the positivity constraint (7), (6), and the trading capacity constraint (8). After forming the Hamiltonian [see Seierstad and Sydsæter (1977)], we find that the first-order sufficient conditions of the market maker's problem are

$$\begin{aligned}
r\lambda_I(t) &= \eta_I(t) + \dot{\lambda_I}(t),\\
p(t) &= \lambda_I(t),\\
\overline{w}(t) &= A(t) - \lambda_I(t) + w_h(t),\\
\overline{w}(t) &= -B(t) + \lambda_I(t) + w_\ell(t),
\end{aligned}$$

where $\lambda_I(t)$ is the co-state variable associated with the ODE (13), $w_h(t)$ and $w_\ell(t)$ are the Lagrange multipliers of the positivity constraints (7) and (6), and $\overline{w}(t)$ is the Lagrange

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

multiplier of the trading capacity constraint (8). We also have the complementary slackness conditions,

$$\eta_I(t) \geq 0 \text{ and } \eta_I(t) I(t) = 0,$$
$$w_h(t) \geq 0 \text{ and } w_h(t) u_h(t) = 0,$$
$$w_\ell(t) \geq 0 \text{ and } w_\ell(t) u_\ell(t) = 0;$$

the jump condition,

$$\lambda_I(t^+) - \lambda_I(t) \leq 0 \quad \text{if } I(t) = 0;$$

and the transversality condition that $\lambda_I(t) e^{-rt}$ goes to zero as t goes to infinity. Direct calculation shows that the equilibrium buffer allocation is a solution of the representative market makers' problem.

To show that high-valuation investors' trading strategy is optimal, we use the Principle of Optimality of dynamic programming. That is, given the price paths $A(t)$ and $B(t)$, it suffices to check that a buyer is worse off if he deviates once from this trading strategy, and behaves according to this strategy thereafter.

During $t \in [0, t_2]$, we need to show that an incoming buyer finds it optimal to purchase the asset immediately rather than waiting for some later time. The utility purchasing immediately at time $t \in [0, t_2]$ is equal to $1/r - A(t)$. The utility of waiting until some time $T > t$ is equal to $e^{-r(T-t)}(1/r - A(T))$. Hence, the net utility of purchasing immediately rather than waiting is

$$\begin{aligned}
&\left[\frac{1}{r} - A(t)\right] - e^{-r(T-t)}\left[\frac{1}{r} - A(T)\right] \\
&\quad = -\int_t^{\min\{T,t_2\}} \frac{d}{dz}\left[\frac{1}{r} - A(z)\right] e^{-r(z-t)}\, dz + \mathbb{I}_{\{T \geq t_2\}} e^{-r(t_2-t)}\left[A(t_2^+) - A(t_2^-)\right] \\
&\quad - \int_{t_2}^{\max\{T,t_2\}} \frac{d}{dz}\left[\frac{1}{r} - A(z)\right] e^{-r(z-t)}\, dz \\
&\quad = \int_t^T \left[-rA(z) + 1 + \dot{A}(z)\right] e^{-r(z-t)}\, dz + \mathbb{I}_{\{T \geq t_2\}} e^{-r(t_2-t)}\left[A(t_2^+) - A(t_2^-)\right] > 0
\end{aligned}$$

because $A(t_2^+) - A(t_2) \geq 0$ and because, from differentiating (34)–(36), it follows immediately that $rA(t) \leq 1 + \dot{A}(t)$.

We also need to verify that a high-valuation owner finds it optimal to hold on to his asset and receive the utility $1/r$. Deviating once from this strategy means immediately selling at price $B(t)$ and purchasing again at some $T > t$. Therefore, the net utility of holding on to the asset rather than selling is

$$\begin{aligned}
&\frac{1}{r} - B(t) - e^{-r(T-t)}\left[\frac{1}{r} - A(T)\right] \\
&\quad = A(t) - B(t) + \left[\frac{1}{r} - A(t)\right] - e^{-r(T-t)}\left[\frac{1}{r} - A(T)\right] \geq 0,
\end{aligned}$$

because $A(t) - B(t) \geq 0$ and because of the previous calculation.

The socially allocation is implemented when sellers progressively sell their assets during $[0, t_4]$, and no seller holds assets during $[t_4, \infty)$. This means that a seller must be indifferent

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

between selling at any time during $[0, t_2]$ and $[t_3, t_4]$, prefer to hold onto his asset during $[t_2, t_3]$, and does not finds it optimal to wait until after time t_4 in order to sell. To verify these conditions, note that the value of selling immediately is equal to $B(t)$ whereas the utility of waiting until some $T > t$ is equal to $\int_t^T (1-\delta)e^{-r(z-t)}\,dz + e^{-r(T-t)}B(T)$. Hence the net utility of selling immediately rather than waiting is

$$\begin{aligned} &B(t) - \int_t^T (1-\delta)\, e^{-r(z-t)}\, dz - e^{-r(T-t)} B(T) \\ &\quad = -\int_t^T \left\{ \frac{d}{dz}\left[B(z)e^{-r(z-t)}\right] + (1-\delta)e^{-r(z-t)} \right\} dz \\ &\quad = -\int_t^T \left[-rB(z) + (1-\delta) + \dot{B}(z)\right] e^{-r(z-t)}\, dz \geq 0. \end{aligned}$$

By construction of $B(t)$, the above expression is equal to zero if $T \leq t_4$, and is strictly positive if $T > t_4$. This establishes the optimality of the trading strategy.

Last, one also needs to verify that a low-valuation nonowner never finds it optimal to buy. The value of buying immediately at time t and reselling at some time $T \geq t$ is equal to

$$\begin{aligned} &-A(t) + \int_t^T (1-\delta)e^{-r(z-t)}\, dz + e^{-r(T-t)} B(t) \\ &\quad = B(t) - A(t) - B(t) + \int_t^T (1-\delta)e^{-r(z-t)}\, dz + e^{-r(T-t)} B(t) \leq 0, \end{aligned}$$

because $B(t) - A(t) \leq 0$ and because of the previous calculation.

A.3. PROOF OF PROPOSITION 3

The accounting identities (4) and (5) make it possible to eliminate two state variables: we choose to keep the measure $\mu_{hn}(t)$ of buyers, the measure $\mu_{\ell o}(t)$ of sellers, and market makers' inventory holdings $I(t)$. Following Seierstad and Sydsæter (1977), the Hamiltonian for the planner's problem is

$$\begin{aligned} \mathcal{H}(t) = {} & \mu_h(t) - \mu_{hn}(t) + (1-\delta)\mu_{\ell o}(t) - \lambda_h(t)\left[-u_h(t) + \dot{\mu}_h(t)\right] + \lambda_\ell(t)\left[-u_\ell(t)\right] \\ & + \eta_h(t)\mu_{hn}(t) + \eta_\ell(t)\mu_{\ell o}(t) + \eta_I(t)I(t) + w_h(t)u_h(t) + w_\ell(t)u_\ell(t) \\ & + \overline{w}(t)\left[2K - u_h(t) - u_\ell(t)\right]. \end{aligned}$$

And Seierstad and Sydsæter's first-order sufficient conditions are

$$r\lambda_h(t) = 1 - \eta_h(t) + \dot{\lambda}_h(t), \tag{39}$$

$$r\lambda_\ell(t) = 1 - \delta + \eta_\ell(t) + \dot{\lambda}_\ell(t), \tag{40}$$

$$r\lambda_I(t) = \eta_I(t) + \dot{\lambda}_I(t), \tag{41}$$

$$\overline{w}(t) = \lambda_h(t) - \lambda_I(t) + w_h(t), \tag{42}$$

$$\overline{w}(t) = -\lambda_\ell(t) + \lambda_I(t) + w_\ell(t), \tag{43}$$

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

together with the positivity and complementary-slackness conditions

$$\eta_h(t) \geq 0 \text{ and } \eta_h(t)\mu_{hn}(t) = 0, \tag{44}$$

$$\eta_\ell(t) \geq 0 \text{ and } \eta_\ell(t)\mu_{\ell o}(t) = 0, \tag{45}$$

$$\eta_I(t) \geq 0 \text{ and } \eta_I(t)I(t) = 0, \tag{46}$$

$$w_h(t) \geq 0 \text{ and } w_h(t)u_h(t) = 0, \tag{47}$$

$$w_\ell(t) \geq 0 \text{ and } w_\ell(t)u_\ell(t) = 0, \tag{48}$$

$$\overline{w}(t) \geq 0 \text{ and } \overline{w}(t)\,(2K - u_h(t) - u_\ell(t)) = 0; \tag{49}$$

and the jump conditions[9]

$$\lambda_h(t^+) - \lambda_h(t^-) \geq 0 \text{ if } \mu_{hn}(t) = 0, \tag{50}$$

$$\lambda_\ell(t^+) - \lambda_\ell(t^-) \leq 0 \text{ if } \mu_{\ell o}(t) = 0, \tag{51}$$

$$\lambda_I(t^+) - \lambda_I(t^-) \leq 0 \text{ if } I(t) = 0. \tag{52}$$

Last, we also impose the usual transversality conditions that $\lambda_k(t)e^{-rt}$ goes to zero as t goes to infinity, for all $k \in \{h, \ell, I\}$. We then guess that $\lambda_h(t) = A(t)$, $\lambda_I(t) = B(t)$, $\lambda_\ell(t) = B(t)$, which we verify as follows:

1. For $t \in [0, t_1)$ the first-order sufficient conditions are satisfied with $\eta_\ell(t) = 0$, $\eta_I(t) = 1 - \delta$, $\eta_h(t) = \delta$, and $w_h(t) = w_\ell(t) = \overline{w}(t) = 0$.
2. For $t \in [t_1, t_2)$, the first-order sufficient conditions are satisfied by $\eta_h(t) = 2 - \delta$, $\eta_\ell(t) = \eta_I(t) = 0$, $w_h(t) = w_\ell(t) = 0$, and $\overline{w}(t) = \lambda_I(t) - \lambda_\ell(t)$. To show that $\overline{w}(t)$ is positive, we use the ODEs for $\lambda_\ell(t) = B(t)$ and $\lambda_I(t) = p(t)$ to note that, when $t \in [t_1, t_2)$, $r\overline{w}(t) = -(1 - \delta) + \dot{\overline{w}}(t)$. Direct integration of this ODE implies that, for all $t \in [t_1, t_2)$,

$$\overline{w}(t)e^{-r(t-t_1)} = \overline{w}(t_1) + \int_{t_1}^{t} (1 - \delta)e^{-r(z-t_1)}\,dz \geq 0,$$

because, by construction of Δ^*, $\overline{w}(t_1) = \lambda_I(t_1) - \lambda_\ell(t_1) \geq 0$.
3. For $t \in [t_2, t_3)$, the first-order sufficient conditions are satisfied by $\eta_h(t) = \eta_\ell(t) = \eta_I(t) = 0$, $w_h(t) = 0$, $\overline{w}(t) = \lambda_h(t) - \lambda_I(t) > 0$, and $w_\ell(t) = \lambda_h(t) + \lambda_\ell(t) - 2\lambda_I(t)$. To verify that $w_\ell(t)$ is positive, recall that

$$rw_\ell(t) = 2 - \delta + \dot{w}_\ell(t) \Leftrightarrow \frac{d}{dt}\left[e^{-rt}w_\ell(t)\right] = -(2 - \delta)e^{-rt} < 0$$

and that, by construction, $w_\ell(t_3) = 0$. Thus, $e^{-rt}w_\ell(t)$ is decreasing and equal to zero at $t = t_3$, so $w_\ell(t)$ must be positive during $[t_2, t_3]$.
4. For $t \in [t_3, t_4)$, the first-order sufficient conditions are satisfied by $\eta_h(t) = \eta_\ell(t) = 0$, $\eta_I(t) = 1 - \delta/2 > 0$, $w_h(t) = w_\ell(t) = 0$, and $\overline{w}(t) = (\lambda_h(t) - \lambda_\ell(t))/2 > 0$.
5. For $t \in [t_4, \infty)$, the first-order sufficient conditions are satisfied by $\eta_h(t) = 0$, $\eta_\ell(t) = \delta$, $\eta_I(t) = 1$, $\overline{w}(t) = w_h(t) = w_\ell(t) = 0$.

The last thing to verify is the jump conditions. Note that, by construction of the co-state variables, $\lambda_I(t)$ and $\lambda_\ell(t)$ are continuous, so their jump conditions (51) and (52) are clearly

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

satisfied. On the other hand, $\lambda_h(t)$ is continuous everywhere except at $t = t_2$, where it jumps up, so the jump condition is also satisfied.

Last, we can verify that the planner does not find it optimal to take assets from high-valuation investors, or transfer assets to low-valuation investors. To see this, we just need to add two positive control variables $u_\ell^-(t)$ and $u_h^-(t)$, which respectively represent the positive flows of assets transferred to low-valuation investors and taken from high-valuation investors. Taking first-order conditions, we find that choosing $u_\ell^-(t) = u_h^-(t) = 0$ is optimal if

$$\overline{w}(t) \geq \lambda_\ell(t) - \lambda_I(t),$$
$$\overline{w}(t) \geq \lambda_I(t) - \lambda_h(t).$$

For instance, the first condition says that the shadow value $\overline{w}(t)$ of a unit of capacity has to be greater than the net utility $\lambda_\ell(t) - \lambda_I(t)$ of taking an asset from market makers' inventories and transferring it to a low-valuation investor. These two conditions are clearly satisfied given that $\overline{w}(t) \geq 0$, $\lambda_\ell(t) - \lambda_I(t) \leq 0$, and $\lambda_I(t) - \lambda_h(t) \leq 0$.

A.4. PROOF OF PROPOSITION 4

When $\varepsilon = 0$, the equation $J(\Delta) = 0$ becomes

$$\left(1 - \frac{\delta}{2}\right) e^{-2r\Delta} - (1 - \delta) - \frac{\delta}{2} e^{-rs/K} = 0$$
$$\Leftrightarrow \left(1 - \frac{\delta}{2}\right) e^{-2r\Delta} = (1 - \delta) + \frac{\delta}{2} e^{-rs/K}$$
$$\Leftrightarrow \Delta = -\frac{1}{2r} \log\left(\frac{1 - \delta + \delta/2e^{-rs/K}}{1 - \delta/2}\right) \equiv F(K, r, \delta, s).$$

Clearly $F(\cdot)$ is decreasing in K, increasing in s. For the partial derivative with respect to δ, note that

$$\frac{1 - \delta + \delta/2e^{-rs/K}}{1 - \delta/2} = 1 - \frac{\delta/2}{1 - \delta/2}\left(1 - e^{-rs/K}\right)$$

is a decreasing function of δ, so $F(\cdot)$ is increasing in δ. Finally, consider the partial derivative with respect to r:

$$\operatorname{sign}\left[\frac{\partial F}{\partial r}\right] = \frac{1}{2r^2} \log\left(\frac{1 - \delta + \delta/2e^{-rs/K}}{1 - \delta/2}\right) + \frac{1}{2r} \frac{\delta/2 s/K e^{-rs/K}}{1 - \delta + \delta/2e^{-rs/K}}$$
$$= \log\left(\frac{1 - \delta + \delta/2e^{-z}}{1 - \delta/2}\right) + z \frac{\delta/2e^{-z}}{1 - \delta + \delta/2e^{-z}}, \tag{53}$$

where $z \equiv rs/K$. Now, the function on the right-hand side of (53) is zero at $z = 0$, and it goes to $\log(1 - \delta) - \log(1 - \delta/2) < 0$ as z goes to infinity. In addition, its derivative with

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

respect to z is

$$
\begin{aligned}
&-\frac{\delta/2e^{-z}}{1-\delta+\delta/2e^{-z}}+\frac{\delta/2e^{-z}}{1-\delta+\delta/2e^{-z}}+z\frac{d}{dz}\left[\frac{\delta/2e^{-z}}{1-\delta+\delta/2e^{-z}}\right]\\
&\quad=z\frac{d}{dz}\left[1-\frac{1-\delta}{1-\delta+\delta/2e^{-z}}\right]\\
&\quad=-(1-\delta)z\frac{d}{dz}\left[\frac{1}{1-\delta+\delta/2e^{-z}}\right]<0.
\end{aligned}
$$

Therefore, the function on the right-hand side of (53) is negative, and so is the partial derivative of $F(\cdot)$ with respect to r.

A.5. PROOF OF PROPOSITION 5

Recall that $I(t_2)=2KF(K,r,\delta,s)$, so that

$$
\begin{aligned}
\frac{\partial I(t_2)}{\partial K}&=F+K\frac{\partial F}{\partial K}=F-K\frac{1}{2r}\frac{\delta/2(rs)/K^2e^{-rs/K}}{1-\delta+\delta/2e^{-rs/K}}\\
&=-\frac{1}{2r}\log\left(\frac{1-\delta+\delta/2e^{-z}}{1-\delta/2}\right)-z\frac{1}{2r}\frac{\delta/2e^{-z}}{1-\delta+\delta/2z},
\end{aligned}
$$

where $z\equiv rs/K$. Proceeding as in equation (53) with the derivative of $F(\cdot)$ with respect to r, one then shows that this last term is positive.

A.6. PROOF OF PROPOSITIONS 6, 7, AND 8

We first prove that t_4 is decreasing in K. We start from equation (20):

$$
t_4=t_2+\frac{s}{K}-\Delta=t_2+\frac{s}{K}-F(K,r,s,\delta).
$$

Taking the derivative with respect to K, we find

$$
\begin{aligned}
\frac{\partial t_4}{\partial K}&=-\frac{s}{K^2}-\frac{\partial F}{\partial K}\\
&=-\frac{s}{K^2}+\frac{1}{2r}\frac{\delta/2e^{-rs/K}(rs/K^2)}{1-\delta+\delta/2e^{-rsK}}\\
&=\frac{s}{K^2}\left[-1+\frac{1}{2}\frac{\delta/2e^{-rs/K}}{1-\delta+\delta/2e^{-rs/K}}\right]<0,
\end{aligned}
$$

where the second equality follows from differentiating the explicit solution for $F(K,r,s,\delta)$ that we derived in the proof of Proposition 4. This proves Proposition 8.

Turning to Proposition 7 we note that the bid price depends only on capacity through the terminal condition $B(t_4)=1/r$ but not on the ODE $rB(t)=1-\delta+\dot{B}(t)$. Consider then two levels $K<\hat{K}$ of capacity, with the corresponding recovery times $t_4>\hat{t}_4$ and bid prices $B(t)$ and $\hat{B}(t)$. During the time interval $[0,\hat{t}_4]$, the bid prices $B(t)$ and $\hat{B}(t)$ solve

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press

the same ODE but with different terminal conditions:

$$B(\hat{t}_4) < B(t_4) = \frac{1}{r} = \hat{B}(\hat{t}_4).$$

Because two solutions of a given ODE for two different initial conditions never cross, and because $B(\hat{t}_4) < \hat{B}(t_4)$, we must have that $B(t) < \hat{B}(t)$ for all $t < \hat{t}_4$. Clearly, this inequality also holds during $[\hat{t}_4, t_4]$, because during that interval $\hat{B}(t) = 1/r$.

Finally, we prove Proposition 6. During $[t_1, t_2]$, $\sigma(t) = 2(1-\delta)/r(1-e^{-r(t-t_1)})$, which is a decreasing function of t_1. Because $t_1 = t_2 - \Delta$ and because Δ is a decreasing function of K, it follows that t_1 is an increasing function of K, so that $\sigma(t)$ is a decreasing function of K. During $[t_2, t_4]$, $\sigma(t) = \delta/r(1-e^{-r(t_4-t)})$, which is an increasing function of t_4 and therefore a decreasing function of K.

https://doi.org/10.1017/S1365100510000714 Published online by Cambridge University Press