ISSN 2279-9362

Collegio Carlo Alberto

Equilibrium price of immediacy and infrequent trade

Riccardo Giacomelli
Elisa Luciano

No. 221
November 2011 (Revised, October 2013)

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Equilibrium price of immediacy and infrequent trade

Elisa Luciano[*], Riccardo Giacomelli [†]

October 24, 2013

Abstract

The paper studies the equilibrium value of bid-ask spreads and time-to-trade in a continuous-time, intermediated financial market. The endogenous spreads are the price at which brokers are willing to offer immediacy. They include physical trading costs. Traders intervene optimally, when the portfolio mix reaches endogenously determined barriers. Spreads and times between successive trades are increasing with the difference in agents risk attitudes. They react asymmetrically to an increase in the difference of risk aversions, while they are symmetric in trading costs. We detect a bias towards cash. Optimal trade is drastically reduced when costs increase, so as to preserve the investors welfare. Random switches to a competitive market, to be interpreted as outside options, drastically reduce bid-ask fees.

It is well known that trading behavior in financial markets is affected by its costs. In a competitive market, they are exogenous trading costs. A non-exhaustive list includes participation costs, such as infrastructure or access costs, information, search and execution costs, including taxes. In a centralized market, traders interact through an intermediary, who faces exogenous costs and, by standing ready to absorb any order from the rest of the market, provides the service of immediacy, or immediate liquidity. He is expected to charge a fee for this service, on top of being reimbursed of the trading costs he absorbs. His bid-ask spread - which represents the overall *transaction cost to investors* - will include both *trading costs*[1] *and the price of immediacy.*

[*] Università di Torino, ICER and Collegio Carlo Alberto, Italy. Email: luciano at econ.unito.it. Elisa Luciano is indebted with Luca Sitzia for computational assistance and thanks Ricardo Caballero, Bernard Dumas, Alessandro Gavazza, Marco Pagano and Larry Samuelson for discussions and suggestions. The comments of Suleyman Basak, Roberto Marfè, Franck Moraux, two suggestions and seminar participants at Rennes 1 University and Collegio Carlo Alberto are gratefully acknowledged. Elisa Luciano thanks two anonymous referees for comments on a solo version of the paper. Any errors are ours.

[†] Università di Torino, Ialy. Email: riccardo.giacomelli@unito.it

[1] Other Authors call “transaction costs” the physical costs, while we call them “trading” or “physical” costs and use the terminology “transaction costs” for the difference between the bid or ask price and its fundamental value.

The way in which trading costs affect equilibrium asset prices in the competitive case, as well as the way in which they interact with the price for immediacy in the intermediated case, is not easy to assess. The reason is that costs go hand in hand with *infrequent trade*, as opposite to the standard continuous trading of frictionless models.

Decentralized models with symmetric information have successfully addressed general equilibrium *asset pricing and trade frequency* in the presence of trading costs, both when investors have the same risk aversion (Vayanos (1998), Lo, Mamaysky and Wang (2004)), and when they do not (Buss and Dumas (2012)). They have investigated the effect of trading costs on prices and turnover. Since trading is competitive, agents simply share *exogenous trading costs.* The sharing rule is endogenous. In Vayanos' overlapping-generation model, costs have a small effect on prices, while the trading frequency is dramatically reduced with respect to a frictionless situation. Investors can refrain from trade even for decades. Lo *et al.* point at a more significant effect of costs on equilibrium prices. Buss and Dumas go even further. They use the assessed effect on prices to produce a cost-adjusted CAPM and to explain some empirical asset pricing puzzles. In terms of trade, the last two papers get smaller times to next trade, since investors have a so-called high-frequency motive to trade, given by an infinite-variation fluctuation in dividends.

Recent models of *centralized* trading instead provide *endogenous* bid-ask spreads but explain it through asymmetric information. These models have concentrated mostly on a specific source of costs, namely search costs, when there is the possibility of trading both in a decentralized and centralized way (Duffie, Gârleanu and Pedersen (2005, 2007)).

This paper aims at filling a gap in the literature, by focusing on *centralized trading with symmetric information. It aims at explaining both the level of endogenous spreads and the amount of endogenous, infrequent trading in general equilibrium.* Our starting point is the case in which there are no physical trading costs, but still the intermediary deserves a fee for the service of immediacy. We study first a situation in which investors must trade through the intermediary, then a situation in which they have the "outside option" of waiting and trade at no cost in a decentralized market. This permits to understand how much equilibrium bid-ask spread, but also trade frequency, are affected by exogenous components, such as physical costs, and competition. Up to the best of our knowledge, this is the first model which endogenizes bid-ask spreads with symmetric information. With respect to decentralized trade, we can split the impact of exogenous trading costs and intermediary services on spreads.

The paper is expected to enhance the comprehension of the price for intermediation and trade impact of strategic brokers behavior. It aims at doing so with respect to the partial equilibrium models of investor's behavior in the presence of transaction costs, such as Constantinides (1986) - which take those *costs as exogenous* - and with respect to the traditional microstructure literature, as exemplified by the seminal models in intermediaries' pricing, such as Stoll (1978), Ho and Stoll (1981) - which takes the frequency of *trade as exogenous.*

In order to study equilibrium bid-ask spreads we go back to the simplest framework for investors' choices in continuous-time stochastic economies, characterized by a risky and a riskless asset, together with infinitely lived, power utility agents. We assume that a representative investor faces a single broker, or specialist, who sets the spreads. We verify that the equilibrium conditions have no solution if the bid-ask spread is null, unless the investor is risk neutral. We show numerically that, if the risk aversion of the agents is diverse, with brokers less risk averse than investors, an equilibrium exists. Spreads and the time to next trade are increasing in the difference in risk aversion, while welfare loss for the investor is not. Bid and ask prices are very sensitive to risk-aversion differences: the impact of the first on the second is one order of magnitude bigger. Also, spreads do not react symmetrically to discrepancies in risk attitudes and generate a bias towards cash. We present our model and results first for the case in which there are no trading costs, then in the presence of trading costs. Opposite to differences in risk aversion, trading costs generate symmetric effects on bid-ask prices and barriers. Last but not least, we extend to the case in which investors can choose either to trade with the specialist at his bid-ask fee or to wait until another investor, with whom they can trade at no cost, submits an order to the market. The second situation, in which investors have an outside-option driven by a regime-switch, provides much smaller fees, as expected.

The outline of the paper is as follows. Section 1 sets up the model without trading costs. Section 2 studies the optimization conditions for the two types of agents (investors and specialist). Section 3 defines equilibrium and studies its features. Section 4 rules out the existence of equilibrium in which spreads vanish. Section 5 provides numerical examples of equilibrium and studies spreads, trading policy, transaction frequency, welfare implications, as well as their sensitivity to the specialist's risk aversion, in comparison to partial equilibrium models. Section 6 covers the case in which exogenous trading costs exist too.. Section 7 studies the outside-option case and its implication for equilibrium spreads and trade. Section 8 summarizes and outlines further research.

1 Model set up

This section specifies the objective of the agents, the *admissible* transaction costs and *admissible* dynamics of traded assets. We consider a continuous-time stochastic economy in which two assets are traded: a riskless and a risky one. The interest rate r on the riskless asset is not determined endogenously. The pre-bid, pre-ask price of the risky asset - its fundamental value, which describes its dividends - is a geometric Brownian motion with parameters α and σ. Two agents populate our economy: a representative investor and a specialist.

The *investor* maximizes the expected utility of his terminal wealth, $\mathbb{E}U(W(T))$. He has an infinite-horizon power utility, $U(W) = W^\gamma/\gamma$. Unless otherwise specified, we assume that he his risk averse and non-myopic: $\gamma < 1, \gamma \neq 0$. His objective is

$$\lim_{T\to\infty} \sup \mathbb{E}U(W(T)) = \lim_{T\to\infty} \sup \mathbb{E}\left[W(T)^\gamma/\gamma\right] \tag{1}$$

The *admissible transaction costs* are proportional to the value of trade. For each dollar value of risky security he trades, the investor receives a bid price s and pays an ask price $1/q$, which will be constrained to be respectively smaller and greater than - or at most equal to - one: $s, q \,\epsilon\, (0, 1]$. We will call the differences $1-s, 1/q-1$ the *transaction* costs, in order to distinguish them from the actual *trading* costs, which impinge on the broker only and will be introduced in section 6 only. The proportionality constants s and q will be determined in equilibrium as a function of all the exogenous variables. Since we will search for a *stationary* equilibrium, s and q will be constant over time.

The investor takes as given the transaction costs, as well as the risk-return features of the risky asset. Let $x(t)$ and $y(t)$ be the (fundamental) values of his riskless and risky position[2]. His final wealth is their liquidation value, i.e. $W(T) = x(T) + sy(T)$.

The (partial equilibrium) investor's optimization problem has been solved by Dumas and Luciano (1991) for the case of non-infinitesimal spreads and by Gerhold, Guasoni, Muhle-Karbe and Schachermayer (2011) for the case of infinitesimal spreads. It is shown in both papers that - if z is the standard Brownian motion which drives y - there exist two increasing processes L and U which make the value of the investor's assets evolve according to

$$\begin{cases} dx(t) = rx(t)dt + sdU(t) - dL(t) \\ dy(t) = \alpha y(t)dt + \sigma y(t)dz(t) + qdL(t) - dU(t) \end{cases} \tag{2}$$

The processes L and U increase only when $\theta \doteq y/x$, the ratio of risky to riskless asset in portfolio, reaches respectively a lower and an upper barrier, which we denote as l and u. Their changes are the local time of the stochastic process θ at the lower and upper barrier. As a consequence, there is no exact notion of "trade size", since the adjustment does not occur in discrete amounts.

In most of what follows, for the sake of simplicity, we restrict our formulas to parameter combinations which make both barriers positive[3], i.e. $0 < l < u$. To this end, we restrict the parameters so that the optimal asset holdings would be positive in the absence of transaction costs:

$$0 < \frac{\alpha - r}{(1-\gamma)\sigma^2} < 1 \tag{3}$$

We know that asset holdings with bid-ask spreads include the optimal holdings in the corresponding frictionless market, $l < \theta^* < u$, where the optimal ratio θ^* is the standard Merton's one:[4]

$$\left(\frac{y}{x}\right)^* = \theta^* = \frac{\alpha - r}{(1-\gamma)\sigma^2 - \alpha + r}$$

[2] Later on y will be called also the pre-spread price: indeed, we do not need to distinguish prices and values.

[3] The computations for the other cases, which were used for the numerical implementations, can be obtained from the Author upon request.

[4] In the absence of costs (and intermediaries) not only individuals would keep their asset ratio at θ^*, but, as demonstrated by He and Leland (1993), a geometric Brownian motion would be the equilibrium asset process for power utility investors.

With trading frictions, the ratio is not kept at θ^*, but the high trading frequency needs imposed by the infinite variation of the underlying fundamental price y remains.

The *specialist* pays to the investor the returns on the risky asset[5] and stands ready to absorb all the transactions required by the investor. He charges a bid and an ask price for this, i.e. he sets s and q. Both the risky and riskless asset are in zero net supply, so that demand equals supply by definition (this will be useful in equilibrium). If x_s and y_s are the specialist's asset holdings, this implies $x_s = -x$, $y_s = -y$ and the ratio $\theta_s = y_s/x_s$ is the opposite of the consumer one, namely $\theta_s = -\theta$. The specialist is a power-utility agent which aims at maximizing the expected utility U_s of his final wealth, when the horizon becomes infinite:[6]

$$\lim_{T\to\infty} \sup \mathbb{E} U_s(W(T)) = \lim_{T\to\infty} \sup \mathbb{E}\left[W(T)^{\gamma'}/\gamma'\right] \tag{4}$$

If not specified otherwise, we also assume that he is risk-averse, $1-\gamma' > 0$. For the time being we assume that he does not incur trading costs. This means that the dynamics of his assets is

$$\begin{cases} dx_s = rx_s dt - sdU + dL \\ dy_s = \alpha y_s dt + \sigma y_s dz - qdL + dU \end{cases} \tag{5}$$

while his final wealth is $W_s = x_s + y_s = -x - y$.

2 Optimization

This section briefly reviews the optimality problem of the investor and introduces ex novo the optimality conditions of the specialist. We search for a stationary solution to both problems.

2.1 Optimization for the investor

The optimization problem of the investor is well understood in the literature. Indeed, it is known that, with positive risk aversion, problem (1) under (2) reduces to solving for the function I the ODE

$$(r\gamma - \beta)\, I(\theta) + (\alpha - r) I'(\theta)\,\theta + \sigma^2 I''(\theta)\,\frac{\theta^2}{2} = 0 \tag{6}$$

[5] In this sense, the risky asset can be interpreted as in Buss and Dumas (2012): it entitles the investor to receive his risky endowment.

[6] We rule out constraints on his wealth. In particular, we rule out the possibility of default of the intermediary.

- with $\beta \in \mathbb{R}$ - under the value-matching and smooth-pasting BCs, namely

$$\begin{cases} lI'(l) = \gamma I(l)\varepsilon_l \\ uI'(u) = \gamma I(u)\varepsilon_u \\ lI''(l) = (\gamma - 1)\, I'(l)\varepsilon_l \\ uI''(u) = (\gamma - 1)\, I'(u)\varepsilon_u \end{cases}$$

where we have used the shortcut notation

$$\varepsilon_l \doteq \frac{l}{l+q} \tag{7}$$

$$\varepsilon_u \doteq \frac{us}{1+us} \tag{8}$$

The function I provides us with the value function of the problem,

$$\lim_{T\to\infty} K(x,y,t;T) \doteq \lim_{T\to\infty} \sup \mathbb{E}\left[W(T)^\gamma/\gamma\right]$$

if there exists a constant β - an artificial discount rate - which makes K itself, once discounted, finite and stationary. Formally, we need β such that

$$\begin{aligned} J(x,y,t;T) &= e^{-\beta(T-t)}K(x,y,t;T) \\ \lim_{T\to\infty} J(x,y,t;T) &= J(x,y) \end{aligned}$$

and, given the homotheticity of the utility function, we assume $J(x,y) = x^\gamma I(\theta)$.

It has also been shown that a solution technique for the above problem consists of three steps. The steps - which are described in Appendix A - turn the investor's problem into an algebraic equation in the unknown δ. Having defined

$$m \doteq (\alpha - r)/\sigma^2 - 1/2, \tag{9}$$

$$\nu \doteq \frac{\sqrt{\left|\left(\alpha - r - \sigma^2/2\right)^2 - 2\delta\sigma^2\right|}}{\sigma^2} \tag{10}$$

and

$$\delta \doteq r\gamma - \beta,$$

the algebraic equation is

$$a(l,q)b(u,s) - c(u,s)d(l,q) = 0 \tag{11}$$

where - using si and co to denote the trigonometric sines and cosines[7] - the

[7] There is also a case where the sines and cosines have to be interpreted as hyperbolic ones, and slight differences in signs occur. The type of solution depends on whether, having defined

$$\delta_{c.} \doteq \frac{\left(\alpha - r - \sigma^2/2\right)^2}{2\sigma^2}, \tag{12}$$

we have $\delta > (<)\delta_c$. (see Appendix A).

expressions for a, b, c, d are

$$\begin{aligned}
a(l,q) &= (-m-\gamma\epsilon_l)\, si(\nu\ln(l)) + \nu co(\nu\ln(l)) \\
b(u,s) &= (-m-\gamma\epsilon_u)\, co(\nu\ln(u)) - \nu si(\nu\ln(u)) \\
c(u,s) &= (-m-\gamma\epsilon_u)\, si(\nu\ln(u)) + \nu co(\nu\ln(u)) \\
d(l,q) &= (-m-\gamma\epsilon_l)\, co(\nu\ln(l)) - \nu si(\nu\ln(l))
\end{aligned} \tag{13}$$

The solutions for δ which are acceptable are the ones which make ε_l and ε_u real. For the case of negative (positive) γ, a straightforward computation shows that this is the case as long as $\delta \leq (\geq)\, \delta^*$, where

$$\delta^* \doteq \frac{\gamma(\alpha-r)^2}{2(\gamma-1)\sigma^2}. \tag{14}$$

2.2 Optimization for the specialist

The specialist aims at maximizing his utility from final wealth, and we let his horizon diverge, while searching for a stationary solution. However, his instruments are not the trading barriers l and u, but the trading costs s and q. The specialist's problem is subject to the standard value-matching conditions, when the processes L and U are different from zero. The FOCs with respect to l and u which provide the smooth-pasting conditions for the investor though must be substituted by optimality conditions with respect to s and q. It can be shown that the value function cannot - and need not - be maximized with respect to s, q on the whole domain, but at most for specific choices of θ. The natural choices are $\theta = l$ and $\theta = u$, since trade occurs at those levels only. Using the traditional approach to smooth pasting, we set the derivatives of the value function equal to zero with respect to s and q at $\theta = l$ and $\theta = u$. Let K_s be the specialist's value function, i.e.

$$\lim_{T\to\infty} K_s(x_s, y_s, t; T) = \lim_{T\to\infty} \sup \mathbb{E}\left[W_s(T)^{\gamma'} / \gamma' \right]$$

It is easy to show, as in the investor's case, that, if we aim at a *stationary* value function, we must discount K_s at a rate $\beta' \doteq r\gamma' - \delta'$. We can define the discounted value function

$$J_s(x_s, y_s, t; T) = e^{-\beta'(T-t)} K_s(x_s, y_s, t; T)$$

and assume that it has a stationary limit:

$$\lim_{T\to\infty} J_s(x_s, y_s, t; T) = J_s(x_s, y_s) = x^{-\gamma'} I_s(\theta).$$

We end up with the following differential equation for I_s :

$$\left(r\gamma' - \beta'\right) I_s(\theta) + (\alpha - r) I_s'(\theta)\, \theta + \sigma^2 I_s''(\theta)\, \frac{\theta^2}{2} = 0 \tag{15}$$

whose solution is of the type

$$I_s = \begin{cases} \theta^{-m}\left[A' si(\nu' \ln(\theta)) + B' co(\nu' \ln(\theta))\right] \\ \mathcal{A}'\theta^{x_1'} + \mathcal{B}'\theta^{x_2'} \end{cases} \tag{16}$$

with $x'_{1,2} = m \pm \nu'$

$$\nu' \doteq \frac{\sqrt{\left|\left(\alpha - r - \sigma^2/2\right)^2 - 2\delta'\sigma^2\right|}}{\sigma^2}$$

The value-matching conditions impose continuity of the value function at the trading points. Indeed, the investor chooses a trading policy which requires his counterpart to trade so as to stay at the boundary of the trading region too. We have:

$$\begin{cases} lI_s\ '(l) = \gamma' I_s(l)\varepsilon_l \\ uI_s'(u) = \gamma' I_s(u)\varepsilon_u \end{cases} \tag{17}$$

where the ε are the ones defined above (and decided by the investor). As in the investor's case, these value matching conditions imply that the constant δ' satisfies

$$a'(l,q)b'(u,s) - c'(u,s)d'(l,q) = 0 \tag{18}$$

where[8]

$$\begin{aligned} a'(l,q) &= (-m - \gamma'\epsilon_l)\, si(\nu' \ln(l)) + \nu' co(\nu' \ln(l)) \\ b'(u,s) &= (-m - \gamma'\epsilon_u)\, co(\nu' \ln(u)) - \nu' si(\nu' \ln(u)) \\ c'(u,s) &= (-m - \gamma'\epsilon_u)\, si(\nu' \ln(u)) + \nu' co(\nu' \ln(u)) \\ d'(l,q) &= (-m - \gamma'\epsilon_l)\, co(\nu' \ln(l)) - \nu' si(\nu' \ln(l)) \end{aligned}$$

In order to take the derivatives of the value function with respect to the specialist' s choice variables, recognize that the bid price s applies at the upper barrier u only, while the ask price $1/q$ applies at the lower barrier l only. As a consequence, the derivatives to be equated to zero are with respect to q at l and with respect to s at u. In taking these derivatives, the broker considers the investor's reaction to his choice of the spreads.[9]

The optimality conditions of the broker are obtained from (17), differentiating with respect to q and s, i.e. computing

$$\begin{cases} \dfrac{d}{dq}\left[-\gamma' I_s\ (l) + (q+l)\, I_s\ '(l)\right] = 0 \\ \dfrac{d}{ds}\left[(1+us)\, I_s\ '(u) - \gamma' I_s(u)s\right] = 0 \end{cases}$$

[8] Here too we report the trigonometric case only. In the investor's case the equation for δ incorporated both the value-matching and smooth-pasting condition, since the ε were determined as in Appendix A. The equation for δ' incorporates the value-matching conditions only, since the ε come from the investors' problem.

[9] It can be demonstrated that an equilibrium in which dealers do not take the reaction of their counterpart into consideration does not exist. The reaction is evaluated in terms of barriers, not in terms of traded quantities, since we know that the investor trades so as to stay along the barriers of the no-transaction cone. The only investor's reaction is in terms of the level, or barrier, not in terms of quantity of intervention, or amount of trade.

which gives the “modified” smooth-pasting conditions

$$\begin{cases} \dfrac{\partial l}{\partial q}\left[\left(1-\gamma'\right) I_s{}'(l)+(q+l)\, I_s{}''(l)\right]+I_s{}'(l)=0 \\ \dfrac{\partial u}{\partial s}\left[(1+us)\, I_s{}''(u)+\left(1-\gamma'\right) sI_s'(u)\right]-\gamma' I_s(u)+uI_s(u)=0 \end{cases} \tag{19}$$

In the last system we have the derivatives of the boundaries with respect to the costs, $\frac{\partial l}{\partial q}, \frac{\partial u}{\partial s}$, which must be obtained from the investor’s problem solution. Appendix B shows that, if we compute appropriately these derivatives and substitute for (17) and (19) into the ODE, we get the following algebraic equations, which synthesize the value-matching and “modified” smooth-pasting conditions for the specialist:

$$\delta' + \varepsilon_l(\alpha - r)\gamma^{'} - \tfrac{\sigma^2}{2}\gamma'\varepsilon_l^2 \left[\tfrac{1}{\frac{\partial l}{\partial q}} + 1 - \gamma'\right] = 0 \tag{20}$$

$$\delta' + \varepsilon_u(\alpha - r)\gamma^{'} + \tfrac{\sigma^2}{2}\gamma^{'}\varepsilon_u^2 \left[\tfrac{u}{s}\tfrac{1-\varepsilon_u}{\varepsilon_u \frac{\partial u}{\partial s}} - 1 + \gamma^{'}\right] = 0 \tag{21}$$

where $\frac{\partial l}{\partial q}$ and $\frac{\partial u}{\partial s}$ are given in Appendix B.

3 Equilibrium

This section defines an equilibrium for the previous economy and comments on the properties of its prices and quantities.

An *equilibrium* in the previous market is a quadruple $\left(\delta, \delta', s, q\right)$, with $s, q \in (0,1]^2$, such that

- the investor’s maximization problem is solved
- the specialist’s one is solved too
- and the barriers l and u are real: $\delta \leq (\geq)\delta^*$ if $\gamma < (>)0$.

Since, by definition, the specialist absorbs any trading need of the investor, we do not need to worry about matching demand and supply of the risky and riskless asset. No market clearing condition is needed, since we are working with a state variable, the investor risky to riskless ratio $\theta = y/x$, which is equal to the opposite of the corresponding ratio for the specialist $\theta_s = -y_s/x_s$. Market clearing is embedded into the choice of the state variable. Overall, an equilibrium requires that the four algebraic equations (11), (18), (20), (21) - which we report here for the sake of convenience - be solved at the same time[10] with $s, q \in (0,1]^2, \delta \leq (\geq)\delta^*$ if $\gamma < (>)0$.

[10] For given $\alpha - r, \sigma^2, \gamma, \gamma'$, the investors’ problem is solved once δ is found, while the specialist’s one is solved once δ', s, q are.

$$\left\{\begin{array}{c} ab - cd = 0 \\ a'b' - c'd' = 0 \\ \delta' + \varepsilon_l(\alpha - r)\gamma^{'} - \frac{\sigma^2}{2}\gamma'\varepsilon_l^2\left[\frac{1}{\frac{\partial l}{\partial q}} + 1 - \gamma'\right] = 0 \\ \delta' + \varepsilon_u(\alpha - r)\gamma^{'} + \frac{\sigma^2}{2}\gamma^{'}\varepsilon_u^2\left[\frac{u}{s}\frac{1-\varepsilon_u}{\varepsilon_u\frac{\partial u}{\partial s}} - 1 + \gamma^{'}\right] = 0 \end{array}\right. \tag{22}$$

Equilibrium prices, quantities and trade are as follows.

3.1 Prices

The procedure we follow consists in verifying that the pre bid, pre-ask geometric Brownian motion[11] price specified above is indeed a *fundamental value.*[12] We know from He and Leland (1993) that it is the equilibrium asset process for the corresponding economy without intermediaries and transaction costs. In an intermediated market, investors sell at a constant discount on it, as commanded by the bid price s, and buy at a surcharge on it, given by the ask price $1/q$. The fundamental value is never observed as a trading price, while sy and y/q are. They can be observed only when trade occurs, though. There are two different trading prices. When trade occurs because the investor reaches his upper barrier, and needs to sell the risky asset, the cum-bid price sy is the *observed trading price*; when trade occurs at the lower investor's barrier, the cum-ask price y/q is the observed trading price. Both prices are reduced (substantially reduced, as we will see in numerical examples) because of transaction costs s and q, even in the absence of trading costs. This is in the spirit of Amihud and Mendelson (1986). The ensuing bid-ask spread (per unit value of the underlying) $1/q - s$ is going to represent the *equilibrium price of immediacy.*

In the traditional microstructure literature the bid and ask prices usually depend on the level of inventories. This happens in our case too, since the barriers l and u represent the agents' inventories, and the equilibrium conditions from which s and q are determined involve l and u. Both prices are still decreasing with inventories[13].

[11] In the inventory-based microstructure literature there is a constant fundamental value of the asset, to which cum-spread prices tend to revert. This mean reversion does not exist in our model, since pre-spread prices are geometric Brownian motions, while spreads are constant and time-independent. This makes our model consistent with the lack of mean reversion on broker's prices, as empirically detected, for instance, by Madhavan and Smidt (1991) in equity markets.

[12] We do not have enough structure to determine asset prices via stochastic discount factors. In this sense the process y is not a standard equilibrium price, but a fundamental value or cumulated dividend process. As such it appears in the individuals' budget constraints. Investors are willing to pay for an asset its stream of dividends adjusted for the bid or ask spread (depending on whether they sell or buy). This is similar to Lo et al. (2004).

[13] A main difference between our model and traditional inventory ones is the lack of mean reversion in the inventory level of the broker (l, u). Some inventory-based models do indeed determine a preferred inventory position for the broker, to which he aims at reverting. In our model the broker's inventory, measured by the ratio $\theta_s = \theta$, fluctuates between l and u, and is kept within those barriers because of the optimal policies of investors. There is no optimal portfolio for the broker itself. As a consequence, we do not have problems in matching the

3.2 Quantities

It is known that L and U are the local times of the process y/x at l, u respectively: trade per unit of time is infinitesimal, with infinite total and finite quadratic variation. In this sense, there is no "order size" in the traditional sense of the microstructure literature. However, knowing that the portfolio ratio stays between the barriers and using the properties of local times of regulated Brownian motion, the moments of trade can be computed.

The price and quantity features just listed are consistent with the findings in Buss and Dumas (2012) for a competitive market. Despite the fact that their bid-ask spread is exogenous and time is discrete, since their endowment evolves as a binomial tree, and our risky asset's fundamental value evolves as Brownian motion, in both cases transaction prices and trades have infinite total and finite quadratic variation.

3.3 Trade frequency

The trading policy behind our equilibrium is such that observed trade is not continuous in time, but *infrequent. The frequency of trade will depend on the distance between the barriers* l and u. The closer the barriers, the more frequent trade will be.

It is clear from the equilibrium conditions that spreads and trade will depend on the *risk aversion* of market participants. It is quite intuitive that an equilibrium will exist if the broker is less risk averse than the investor. Pagano and Roell (1989) already proved that brokers trade only with customers more risk averse than themselves.[14] Before investigating this - as well as the spread and trade dependence on the difference in risk aversion between market participants - in the next section we show that the spread can be zero if and only if the specialist is never requested to participate in the market. In turn, we know from optimality conditions without bid-ask spreads that this occurs if the investor is risk-neutral.

4 No bid-ask spread case

We do not expect the broker to provide immediacy for free, by setting the bid and ask price equal to the fundamental price, $s = q = 1$. We can envisage an exception, though: the broker can accept zero fees, if he does not need to intervene. We expect an equilibrium without costs to exist if the investor does

lack of empirical mean reversion in inventories.

[14] In Pagano and Roell's set up, brokers set the bid-ask price competitively, by equating the utility they get with and without operating as brokers. Investors equate the utility they get when selling (buying) in a brokers' market with the one they get when selling (buying) in a competitive market, i.e. an auction or limit-order one. When customers are more risk averse than brokers, the possibility of trading depends also on the spread which would prevail on a competitive market and on the probability of finding a counterpart in it. When trade occurs in the brokers' market, the spread magnitude depends on the difference between the risk attitudes of brokers and investors, exactly as in our setting.

not trade. We can verify that this is correct for risk-neutral investors, since we know that - without transaction costs - the no-trade barriers collapse to the optimal Merton's holding, θ^*, and that for risk-neutral investors this holding is -1.

We aim at proving that the system (22) admits no solution when $s = q = 1$, and consequently $l = u = \theta^*$, unless $\gamma = 1$. The proof is as follows. With $s = q = 1$ or no costs, the first two equilibrium conditions in (22) do not determine the rates δ, δ' any more. At $\theta = \theta^* = \frac{\alpha - r}{(1-\gamma)\sigma^2 - \alpha + r}$ the rates of growth are instead the usual

$$\begin{cases} \delta = \frac{(\alpha-r)^2}{2\sigma^2} \frac{\gamma}{1-\gamma} \\ \delta' = \frac{(\alpha-r)^2}{2\sigma^2} \frac{\gamma'}{1-\gamma'} \end{cases}$$

The last two equilibrium equations in (22), once evaluated at $s = u = 1, l = u = \theta^*$, become

$$\begin{cases} \frac{(\alpha-r)^2}{2\sigma^2} \frac{\gamma'}{1-\gamma'} + \frac{(\alpha-r)^2}{(1-\gamma)\sigma^2} \gamma' - \frac{\sigma^2}{2} \gamma' \left(\frac{\alpha-r}{(1-\gamma)\sigma^2} \right)^2 \left[\frac{1}{\frac{\partial l}{\partial q}|_{l=\theta^*, s=q=1}} + 1 - \gamma' \right] = 0 \\ \frac{(\alpha-r)^2}{2\sigma^2} \frac{\gamma'}{1-\gamma'} + \frac{(\alpha-r)^2}{(1-\gamma)\sigma^2} \gamma' + \frac{\sigma^2}{2} \gamma' \left(\frac{\alpha-r}{(1-\gamma)\sigma^2} \right)^2 \left[\frac{\frac{(1-\gamma)\sigma^2 - \alpha + r}{\alpha - r}}{\frac{\partial u}{\partial s}|_{u=\theta^*, s=q=1}} - 1 + \gamma' \right] = 0 \end{cases}$$

They can be solved at the same time if and only if the ratio of the barrier's sensitivity to costs is equal to the optimal riskless to risky holding:

$$-\frac{\frac{\partial u}{\partial s} |_{u=\theta^*, s=q=1}}{\frac{\partial l}{\partial q} |_{l=\theta^*, s=q=1}} = \frac{(1-\gamma)\sigma^2 - \alpha + r}{\alpha - r} > 0$$

It is easy to show that the previous equality - under mild technical conditions on the derivatives of a, b, c, d - is satisfied if and only if $\gamma \to 1$, i.e. the investor is risk neutral. We indeed know that in this case is optimal investment allocation $\theta^* \to -1$, and he does not need an intermediary to optimally balance his wealth. This shows that risk-neutral investors - who do not trade - may be granted zero bid-ask spreads, as intuition would command.

5 Examples

The equilibrium conditions provided above cannot be solved explicitly. We discuss them starting from a base-case, which is calibrated to the pioneering literature in single investor's optimality with transaction costs (Constantinides (1986)). We expect the spreads to be quite bigger than the observed ones, since we have homogenous investors and no outside-option. We are also ready to obtain a frequency of trade low with respect to actual market frequencies, since, on top of the presence of two agents only, in order to keep the model tractable, we disregard some important motives to trade, such as speculative reasons arising from asymmetric information or hedging motives due to incompleteness of the market (also without transaction costs). In this respect, the results of the base-case should be interpreted as those of Buss and Dumas (2012) or Lo *et al.* (2004): there is no attempt to calibrate a specific market.

In section 5.1 we obtain the equilibrium quadruple in the base-case and discuss the resulting bid-ask spread, transaction policy, expected time to next trade and rate of growth of derived utility, in comparison with their partial equilibrium (or investor-only) values. In section 5.2 we discuss the impact of the difference in risk aversions on the results.

5.1 Base case

Starting from the fundamental risk-return base-case in Constantinides (1986), i.e. $\alpha - r = 5\%, \sigma^2 = 4\%$, we assume a coefficient of risk aversion for the investor equal to $1-\gamma = 4$, and a broker's risk aversion slightly smaller: $1-\gamma' = 3.85$.

This section shows - among other things - that

- spreads are one order of magnitude bigger than the (percentage) difference in risk aversion which justifies them, but expected times to next trade are lower than in the corresponding partial equilibrium models. These models were by definition unable to capture the effect of risk-aversion heterogeneity among market participants. By so doing, they overestimated trade inertia, for a given level of costs. The result we obtain reconciles low heterogeneity in risk aversion - which seems to be an empirically relevant phenomenon, see for instance Xiouros and Zapatero (2010) and references therein - with reasonable levels of trade frequency. These are close to weeks or months, not to years or decades as in similarly-calibrated partial-equilibrium models;
- the no-trade region presents a bias toward cash. This bias does not depend on consumption-on-the way. It just depends on the bigger sensitivity of ask prices with respect to risk-aversion difference. This could help in explaining the equity-premium puzzle.

The investor-broker equilibrium is indeed characterized by the quadruple[15]

$$\left(\delta, \delta', s, q\right) = (0.023428, 0.023687, 97.53\%, 68.41\%),$$

with barriers equal to

$$l = 0.301825 < \theta^* < u = 0.480013.$$

since the corresponding no-cost problem has optimal portfolio mix

$$\theta^* = 0.4545$$

[15] For the given parametrization, $\delta_c = 0.01125$, $\delta^* = 0.0234375$. Since both δ and δ' are greater than δ_c, the roots of the algebraic equation corresponding to (6), which is equation (42) in Appendix A - and its equivalent for the broker - are imaginary. The transaction boundaries are real, since $\delta < \delta^*$.

Let us denote with an index p the corresponding partial-equilibrium solutions. Keeping costs at the level provided in general equilibrium, for the sake of comparison, barriers become equal to[16]

$$l_p = 0.1495 < \theta^* < u_p = 0.8243.$$

while the discount rate δ_p becomes 0.0192.

5.1.1 Bid-ask spread

Let us comment on the equilibrium bid/ask spread first. The equilibrium bid price is approximately equal to $s = 97.5\%$ of the pre-bid quote, the ask price is equal to $1/q = 1/68.4\% = 146\%$ of it. The bid-ask spread - or round-trip cost - amounts to $1/q - s = 48.5\%$. With a unique broker and no outside-option, a tiny difference in risk aversion (3.75%) justifies huge costs and a huge *spread* in equilibrium. The latter is *one order of magnitude bigger than the risk aversion (percentage) difference.* This seems to be a very high number, but finds a justification in the facts mentioned at the beginning of section 5. The bid-ask spread is not calibrated to empirically observed values. By using the parameters of the previous transaction-cost, partial-equilibrium literature, we simply aim at stressing how important a subtle difference in risk aversion of market participants can be in terms of price of immediacy. It is very likely to be affected also by the monopoly power of the broker. For this reason, in a later section we weaken his position by introducing outside options. We consider the monopolistic case worth analyzing, because of the sensitivities and asymmetries it unveals, more than because of the absolute level of spreads it entails.

5.1.2 No-trade region

Let us see the effects on the no-trade region. If costs are kept the same between the general and partial equilibrium (in the former being endogenous), we find that the intervention barriers are further apart in the partial-equilibrium than in the equilibrium case:

$$l - l_p = 0.15, u_p - u = 0.34$$

and the no-transaction cone in partial equilibrium incorporates the general equilibrium one:

[insert here figure 1]

This means that partial-equilibrium models are likely to have *overstated the magnitude of no trade*, even though they perfectly captured the trading mechanism. In a general-equlibrium perspective, the investor is less reluctant to trade, since the specialist has forecasted his customer's reaction when fixing

[16] The investor's problem is solved by $\delta = 0.0192 > \delta_c$.

the costs. In terms of optimal overall portfolio mix, as measured by the ratio of risky to total assets, $y/(x+y)$, the equilibrium values are

$$\frac{l}{1+l} = 0.23, \frac{u}{1+u} = 0.32$$

while the partial-equilibrium ones are

$$\frac{l_p}{1+l_p} = 0.13, \frac{u_p}{1+u_p} = 0.45$$

The no-cost optimal mix would be

$$\frac{\theta^*}{1+\theta^*} = 0.31$$

As expected, even in terms of overall portfolio mix, *the barriers are closer to the no-cost situation in the general-equilibrium case.* Both in terms of risky to riskless ratio and overall portfolio mix, the percentage differences between the general and partial equilibrium situation *are one order of magnitude bigger than the risk-aversion difference which justifies them.*

5.1.3 Bias towards cash

The barriers of intervention of the investor are less symmetric with respect to the optimal ratio in the absence of costs, i.e. $\theta^* = 0.45$, than without an intermediary, i.e. in partial equilibrium. This results from the comparison of the barriers

$$\theta^* - l = 0.15, u - \theta^* = 0.03$$

$$\theta^* - l_p = 0.30, u_p - \theta^* = 0.37$$

or from the comparison of the optimal portfolio mix:

$$\frac{\theta^*}{1+\theta^*} - \frac{l}{1+l} = 0.078, \frac{u}{1+u} - \frac{\theta^*}{1+\theta^*} = 0.013$$

$$\frac{l_p}{1+l_p} - \frac{\theta^*}{1+\theta^*} = 0.42, \frac{u_p}{1+u_p} - \frac{\theta^*}{1+\theta^*} = 1.46$$

This permits us to comment on the *bias towards cash* - the riskless asset - which Constantinides found in the partial equilibrium model with consumption. There it was justified by consumption itself, since it vanished with interim consumption, unless the horizon were finite (Liu and Loewenstein (2002)). In our case the bias comes back, even without interim consumption, *since the equilibrium magnitude of costs is not symmetric*: ask spreads $1/q - 1$ are much bigger than bid ones $1 - s$. We are going to argue below that investors would command a *liquidity premium* to switch from an hypothetical market without costs - hypothetical because it would have no trade - to an equilibrium with costs. The presence of a bias towards cash suggests that this liquidity premium should be particularly high.

5.1.4 Trade frequency

The spread and no-trade region features produce trade frequencies which - all others equal - are more realistic than partial equilibrium frequencies. Partial equilibrium models were overestimating the reduction in trade provided by transaction costs. Still, trade is far from being continuous. The frequency of trade can be measured by the expected time that the process θ takes in order to reach either the upper u or the lower barrier l, starting from the optimal mix θ^*. Between l and u, θ has drift $\mu = \alpha - r$ and diffusion σ. Standard results in the theory of the first passage time of a Brownian motion through either an upper or a lower boundary tell us that the expected time we are searching for is

$$t^* = \frac{1}{\frac{\sigma^2}{2} - \mu}\left[\ln\left(\frac{\theta^*}{l}\right) - \frac{1 - \left(\frac{\theta^*}{l}\right)^{1-\frac{2\mu}{\sigma^2}}}{1 - \left(\frac{u}{l}\right)^{1-\frac{2\mu}{\sigma^2}}}\ln\left(\frac{u}{l}\right)\right]$$

In the cum-specialist equilibrium just described, the expected time between transactions is close to 6 months:

$$t^* = 0.508$$

A tiny difference in risk aversion of the participants then makes trade infrequent. The expected time would be

$$t^* \simeq 13$$

years in the corresponding partial-equilibrium case, since the barriers are more distant from the Merton's line.

Partial equilibrium models, all others equal, were *overestimating* the impact not only on the no-trade region, but also on trade infrequency. The expected time between interventions we obtain is *huge* in comparison with the continuous trading of the frictionless literature, but more realistic than the partial equilibrium one. This is in line with what Constantinides aimed at showing, early in the development of the transaction-cost literature, as well as with empirical evidence, largely interpreted, where trading is never a continuum.

Let us compare with the trading frequency obtained in the competitive-equilibrium models of Buss and Dumas (2012) and Lo *et al.* (2004), in which agents split exogenous trading costs. With a similar assumption on the agents' endowment (infinite variation), similar values for the fundamental value of the risky asset (instantaneous return and diffusion) and a much bigger difference in the agents' risk aversion, Buss and Dumas get a mean waiting time between successive transactions which goes up to two years, when the round-trip transaction cost is 20%. They have a trade frequency similar to our with smaller costs, then. Since in their model there is no intermediary extracting a rent from investors, this says that similar trade infrequencies are consistent with different market organizations. In a competitive market, it is achieved by investors further apart in risk aversion, with smaller - but exogenous - costs. Here it is achieved with smaller risk aversion difference and higher costs (due to the rent). Lo *et al.* have an high frequency motive for trade and fixed costs. So, their trade should be boosted by the first motive, kept low by the second. As a result, they have calibrated examples in which - for volatility levels comparable to our choices - the expected time between transactions is close to our. We interpret this result as showing that, as in Buss and Dumas' case, different market settings can provide the same optimal trading frequency. In Lo *et al.*, the trading frequency can reach years, when fixed costs increase. In our case such a high trading frequency would require a much higher difference in risk aversion (see below).

Last, we can compare with partial-equilibrium models with transaction costs and a finite horizon. Our results are in line with the assertion of Liu and Loewenstein (2002), who note that "even small transaction costs lead to dramatic changes in the optimal behavior for an investor: from continuous trading to virtually buy-and-hold strategies. They are less extreme, since in their case costs ranging from 3 to 16%, i.e. in the order of magnitude of s above, together with the same expected return and volatility and similar risk aversion, led to expected transaction times of around 10 to 20 years. The difference is due to the finite horizon, which - all others equal - makes more unlikely that costs can be recouped. As a consequence, the frequency of trade drops even more dramatically than in an infinite-horizon case, where transaction costs can be compensated by the excess return on risky securities. As soon as transaction costs are not infinitesimal, the finite and infinite-time expected transaction frequencies are quite significantly different. In order to achieve a buy and hold strategy in the present setting, while keeping all the other parameters fixed, we should consider a market maker with lower risk aversion, i.e. risk aversion much further from the investor's one. We will indeed see below that, when his risk aversion lowers, and gets further from his counterpart's one, the investor's trade frequency decreases.

5.1.5 Welfare implications

We still need to verify that at least in the base-case welfare - which here is measured by the rate of growth of expected utility - moves in the right direction when going from a non-intermediated market to an intermediated one. To do so, let us comment on the last couple of equilibrium parameters, namely δ and δ'. They indeed determine the rate of growth of the indirect utility of the investor and broker, β and β' respectively[17]. Given that $\beta = r\gamma - \delta$, the higher is δ, the smaller the rate of growth of the corresponding agent. Analogously for β'. Since

$$\delta = 2.34\% > \delta_p = 1.92\%,$$

the investor's rate of growth of expected utility in the current equilibrium, β, is smaller than in the corresponding partial equilibrium. The presence of a (monopolistic) market maker affects this rate in the expected direction.

[17] Recall that, since utility stays bounded when discounted at the rate β, it grows at β when it is not artificially discounted.

Starting from this, we could determine the *liquidity discount* that investors would tolerate, costs being equal, in order to go from a general to a partial equilibrium. This means to determine under which $\alpha - r$ investors see their welfare growth unaffected by the strategic specialist's intervention. Practically, it means to solve for $\alpha - r$ the investor's problem with $\delta_p = 2.34\%$.

We could also determine which *liquidity premium* investors would command in order to keep their welfare unaffected when going from a no-cost equilibrium to an intermediated one. This computation, which would parallel the exam done for partial equilibrium by Constantinides (1986), has the disadvantage that - as we know - the no-cost equilibrium is made by homogenous agents and has no trade. In the base-case studied so far, it would consist in solving for $\alpha - r$ the investor's and specialist's problems. In doing that, it would be necessary to keep the rate of growth of derived utility of the investor at the level it has in the absence of costs, while the broker's rate should be kept fixed at the equilibrium-with-costs level, $\delta' = 2.34\%$.

It has been shown by Constantinides (1986) that the liquidity premium has a second order effect on asset prices, in the sense that the ratio of the liquidity premium - as defined above - to percentage transaction costs is smaller than one; Jang, Koo, Liu and Loewenstein (2007) showed that - if returns are not IID - its effect is of the first order, since the ratio is above one. Both papers drew the conclusion in partial equilibrium. Buss and Dumas (2012) explore the issue in a competitive, general equilibrium whose assumptions are similar to our, apart from the fact that - since there is perfect competition - there is no endogenous price for immediacy. Only exogenous trading costs exist. The liquidity premium they find - after having solved for the pricing kernel - is one order of magnitude smaller than the trading costs, but increasing (and concave) in them. One can conclude from their CAPM that transaction costs are likely to contribute to the explanation of the *equity premium puzzle*, the more so the higher are transaction costs. Since our costs for immediacy are ten times bigger than Buss and Dumas' trading costs, the possibility of having both intermediated and competitive trade could probably raise the liquidity premium so as to push the equity premium in the right direction.

5.2 Sensitivity analysis

This section explores the spread, trade and welfare implications of changing the participants' risk aversion. By decreasing the risk aversion of the broker, or making it further from the investor's one, we find equilibria characterized by lower s and q, which means that bid prices decrease, ask prices and the overall spread and transaction costs increase. Table 1 below gives a comprehensive sensitivity analysis of the equilibrium bid prices, q values, control or no-trade limits, expected time between interventions and growth rates as a function of the broker's risk aversion[18].

[18] A similar analysis could be conducted by varying the volatility parameter σ.

Table 1

$1-\gamma'$	s	q	l	u	$t*$	δ'	δ
3.85	0.975	0.684	0.3018	0.4800	0.508	0.023687	0.023428
3.815	0.946	0.656	0.2869	0.4992	0.973	0.023824	0.023421
3.8	0.935	0.645	0.2811	0.5068	1.176	0.023892	0.023418
3.75	0.900	0.610	0.2644	0.5294	1.839	0.024166	0.023412
3.7	0.869	0.582	0.2530	0.5465	2.385	0.024506	0.023411
3.6	0.841	0.560	0.2435	0.5647	2.973	0.024904	0.02341

The bid-ask *prices* behavior is presented in figure 2 below, as a function of the difference between the agents' risk aversions, $\gamma' - \gamma$:

[insert here figure 2]

Both s and q decrease, at a similar rate. This has an asymmetric impact on trading prices, since the bid price goes down from 97.5% to 84%, while the ask one increases from $1/68.4\% = 146\%$ to $1/56\% = 178\%$. The absolute difference is approximately 13 percentage points in the first case, 32 in the second. The behavior of costs with respect to the difference in risk aversion is apparently counter intuitive. The more distant agents are in risk aversion, i.e. the better risk sharing should work, the higher is the spread. However, we will show in a few lines that the results on barriers, trade frequency and - consequently - the derived utility growth will reconcile this fact with intuition. Welfare - resulting from spreads and optimal trade - moves as risk-sharing commands, even though spreads do not seem to move in the intuitive direction.

Compare now with the microstructure, inventory-based models, such as Ho and Stoll (1981). In most of these models the ask price increases and the bid one decreases with the intermediary's risk aversion. Table 1 shows that in our case the opposite holds: as risk aversion increases, $1/q$ decreases while s increases. This happens exclusively because the *difference* in risk aversion between the two counterparts matters. In our model the ask price $1/q$ decreases and the bid one s increases - thus reducing the bid-ask spread - as the broker's risk aversion goes up and the investor's one remains fixed, i.e. when the difference between their risk aversions goes down (from .4 to .15 in Table 1).

As for *trading barriers*, since lower risk aversion for the broker entails increases in transaction costs, the lower barrier l decreases, while the upper one u goes up. In the plane $x - y$, the cone of no-transactions, characterized by $l < \theta < u$, becomes wider. Investors become more tolerant with respect to discrepancies between their actual asset mix and the optimal, Merton's one, θ^*. In partial equilibrium, this happens as a result of an increase in the investor's risk aversion (see for instance Constantinides (1986)). Here, even if the investor's attitude towards risk does not change, his counterpart's decreased risk aversion makes him more reluctant to trade, since his costs in doing so increase. Figure 3 shows the behavior of the barriers as a function of the difference between the specialist and investor's risk aversion

[insert here figure 3]

By putting together the behavior of the bid-ask spread and the barriers, and recalling that barriers correspond to inventories in the microstructure literature, we observe that not only both the bid and the ask price separately depend on inventories, as was clear from the equilibrium definition, but also the bid-ask spread does. Indeed, going down Table 1, the spread changes and the barriers do. In traditional microstructure models, inventories disappear as determinants of the spread, while being determinants of its components, the bid and ask prices, because of symmetry and linearity assumptions in the demand by investors. O'Hara (1997) already anticipated that independence of the spread from the level of inventories was not very intuitive, and could probably be overcome by relaxing the traditional assumption of a constant fundamental - or pre-spread - value for the underlying good[19]. Our model has no symmetry and linearity assumptions on demand, which is endogenized. More than that, and consistently with O'Hara's intuition, our equilibrium builds on a non-constant fundamental value. Table 1 shows that the bid price s is countermonotonic with respect to the upper trading barrier u: the higher is the broker inventory, the lower is his bid price. The ask price $1/q$ is countermonotonic in the lower barrier l: the higher is the broker inventory, the lower is his ask price. The bid-ask spread $1/q-s$ almost doubles when going from top to bottom in the Table, as O'Hara's suggestion commands.

Figure 3 reports the *optimal holdings without transaction costs* θ^* too. By so doing, it puts into evidence the asymmetry, or *bias toward cash*, when transaction costs increase. Going down the Table, the lower barrier l departs from the optimal ratio without transaction costs, $\theta^* = 0.454\,5$, more than the upper one u: the cone opens up more towards the lower part and people tend to hold more cash than if the barriers opened in a symmetric way. This effect, which we noticed for the base-case, is preserved when costs increase because risk aversions depart. It is due to the interaction of broker and investor, which makes costs on the ask side increase more than costs on the bid side. It is the effect on trade of the greater sensitivity of ask with respect to bid prices. It follows from the sensitivity of s and q with respect to the difference in risk aversions, visualized in Figure 2.

The *frequency of trade* adjusts according to the barriers' movement: the expected time to the first intervention t^* goes up from 6 months to 3 years when risk aversion of the broker decreases. So, in order to obtain a trade frequency of the order of decades we would probably need a very high difference in risk aversion.

The *rates of growth of indirect utility* move too: δ slightly decreases, while δ' increases when risk aversion of the broker decreases (or the two get further apart). This means that $\beta = r\gamma - \delta$ slightly increases. The adjustment of trade - i.e., the opening up of the no-transaction region between l and u - is so large as to make the whole rate of growth of utility go up, even if transaction costs increase.

[19] All others equal, she claims that "the movement of a fixed spread around the true price may no longer be optimal if the price itself is moving".

Not only intervention is rare in time, but such policy is so effective that it may make the whole rate of growth of utility increase even when transaction costs go up[20]. The less a broker is risk averse, i.e. the further in risk aversion he is from his customer, the less his prices will be advantageous for the latter. However, the latter decreases trade so much that his utility's growth rate increases. So, an increase in risk sharing possibilities - because the two market participants are further apart in risk aversion - does not show up in the bid-ask spread. It shows up, as it should, in the welfare of investors.

6 Equilibrium with physical costs

We now focus on an equilibrium in which the specialist imposes the bid-ask spread represented by s, q to his counterpart, but suffers external or physical costs of trading, which he does not pocket. This increases his cash outflow for each unit of risky asset bought from s to $s' > s$, while - for any unit of cash inflow - it modifies the value of the risky-asset sale to $q' > q$. Trading costs are exogenous, as in Lo et al. (2004) or Buss and Dumas (2012); without loss of generality, we assume that the ratio $s'/s, q'/q$ is equal to $k > 1$.

6.1 Model update

Trading costs can be modelled by keeping s and q in the investor's SDEs (2) and inserting $s' > s, q' > q$ in the specialist's ones (5). The broker's final wealth becomes $W_s = x_s + (s' - s)y_s = -x - (s' - s)y$.

It is easy to show that, in order to find an equilibrium with exogenous trading costs on top of endogenous bid-ask spreads, one needs to solve for $\left(\delta, \delta', s, q\right)$ the following equations:

$$\begin{cases} ab - cd = 0 \\ a'b' - c'd' = 0 \\ \delta' + \varepsilon_l'(\alpha - r)\gamma^{'} - \dfrac{\sigma^2}{2}\gamma'\varepsilon_l'^2 \left[\dfrac{1}{\frac{\partial l}{\partial q}} + 1 - \gamma'\right] = 0 \\ \delta' + \varepsilon_u'(\alpha - r)\gamma^{'} + \dfrac{\sigma^2}{2}\gamma^{'}\varepsilon_u'^2 \left[\dfrac{u}{sk}\dfrac{1 - \varepsilon_u}{\varepsilon_u \frac{\partial u}{\partial s}} - 1 + \gamma^{'}\right] = 0 \end{cases} \tag{23}$$

[20] The opposite could happen for the broker: when his risk aversion decreases, β' could decreases. This would mean that, in spite of applying higher costs, he suffers in terms of utility growth, because investors do not visit him very often. Since r is not specified, though, we do not know whether the broker's utility does cumulate at a higher or lower rate β'. It is interesting to notice that the two rates tend to coincide when risk aversions do (i.e., when $1 - \gamma' \rightarrow 4$),as one expects. Recall though that when the spread disappears equilibrium vanishes (since $\gamma \neq 1$).

where we have defined

$$\begin{aligned}\varepsilon_l' &\doteq \frac{l}{l+q'} = \frac{l}{l+qk}\\ \varepsilon_u' &\doteq \frac{us'}{1+us'} = \frac{usk}{1+usk}\end{aligned}$$

6.2 Numerical results

Using the same asset parameters as in the base-case, namely $\alpha - r = 5\%, \sigma^2 = 4\%$, and keeping the investor's risk aversion at $1-\gamma = 4$, as in that case, we explored the equilibrium for a number of possible impacts of external costs k and broker's risk aversion $1-\gamma'$.

Knowing that - without external costs - equilibria with moderate bid-ask spread exist when the risk aversions of the two agents are closer - the investor's one being still bigger than his broker's - we explore here the case in which the two differ by 1%, i.e. $1-\gamma' = 3.96$. As soon as the risk aversion difference is such as to produce moderate transaction costs in the absence of external costs (as it happens when $\gamma' \to \gamma$), we have equilibria also with very high external costs. All others equal, we considered several levels of external costs k. In table 2 we present three of them, for k between 22% (bottom) and 29% (top). We find that s goes from 99.5 to 98.2%, q from 65.5 to 60%.

Table 2

k	s	q	l	u	$t*$	δ'	δ
1.29	0.982	0.597	0.2346	0.5318	2.232	0.023327	0.023212
1.223	0.997	0.626	0.2488	0.5182	1.721	0.023354	0.023245
1.219	0.9953	0.655	0.2688	0.5043	1.211	0.023402	0.023324

These equilibria provide us with new information. They are able to tell us that the price of immediacy - and the width of the no-trade cone - grows when the intermediary suffers external costs - as expected - and how it does. First, the bid-ask spread is *monotonic* in the magnitude of external costs, as expected. In the range examined, the bid-ask spread goes from 69% (top) to 53% (bottom). It is much greater than it is without costs: in Table 1, the spread was already 49% with a risk aversion of 3.85 on the part of the broker, while here it is 53% at the minimum, even though the risk aversions are much closer. Second, also the cone opens up in a *monotonic* way. The lower trading barrier increases, the upper one decreases from top to bottom. Third, the increase is not so much pronounced: even though physical costs range from 22 to 29% of s or q, the bid-ask spread and the barriers are not so far from what they were for pure rent. The spread - as well as the cone - seems to be mostly justified by the market structure, not by physical costs. Fourth, the increase in the spread and the opening of the no-trade region are almost *symmetric* for sales and purchases. From top to bottom, the lower barrier goes from 0.2346 to 0.2688, while the upper one ranges from 0.5318 to 0.5043, so the difference is more or less 0.03 in the first and 0.04 in the second case. Even though the barriers do not have the same sensitive with respect to physical costs, since the absolute change in the lower barrier is more or less the double than the change in the upper one, that difference in sensitivity remains the same when the level of physical costs changes. There is *no asymmetry or bias* towards cash generated by an increase in costs. Only the initial distance from the optimal ratio θ^* (and its sensitivity) is bigger for the lower than for the upper barrier.

The expected time to trade after a re-adjustment increases with respect to the no-cost case, as expected from the spread behavior. Reading now from bottom to top, it ranges from 1.21 to 2.23 years. When costs go up, trade becomes more infrequent. As a result, the investor's δ goes slightly down, his welfare β slightly increases. So, as in the case without costs, the trading policy is so effective that it counterbalances the specialist's pricing policy.

Overall, Table 2 provides further ground for the comparisons with the models of Buss and Dumas or Lo *et al.* which we conducted in the previous section. In our case, including or not physical costs - which are the only ones in the related literature - does not move very much the numerical results and does not change the symmetries or asymmetries, including the bias toward cash and the effectiveness of the trading policy in terms of welfare.

The presence of spreads comonotonic with external trading costs permits to compare with the models of decentralized trading with endogenous spreads mentioned in the introduction, i.e. Duffie et al. (2005). Even though they have asymmetric information and outside options, which are respectively ruled out and do not make sense in our model, Duffie et al. (2005) show that bid-ask spreads are lower if the chance to meet and trade with another agent is easier. In their setting, this typically happens for "big" traders, who are able to contact more counterparts. Such result makes their contribution profoundly different from the traditional information-based literature, which assigns greater spreads to more informed - intuitively, "bigger" - investors. In our setting, Duffie's et al. results can be reproduced by comparing markets with different trading costs. Our cross-market predictions then are similar to their, i.e. give lower spreads when the access to counterparts is easier, for instance because they are "big". In this sense, our cum-broker equilibrium provides an extremely stylized description of OTC markets, certainly poorer than the Duffie et al. one, but with an explicit, motivating role for risk aversion[21]. In addition, since trade frequency is endogenous in our setting, the traders which deserve smaller spreads are the ones which intervene more frequently. This is consistent with them being "big" traders.

[21] The extension to risk-averse market participants in Duffie *et al.* (2007) introduces a role for differential behavior in front of the risky asset. In their case markets participants all have the same risk attitude, but are heterogeneous in background risk, i.e. in the correlation between their endowment and the risky asset. As a consequence, a direct comparison with our difference in risk attitude does not seem to be straightforward.

7 Equilibrium with outside option

7.1 Model

In this section we give investors the outside option to wait and trade in a competitive market, instead of trading with the intermediary. This should enable us to understand how much "competition" is likely to affect equilibrium bid-ask spreads. In order to model the outside option, we assume that over the next instant the market can still be an intermediated one, or investors can find themselves in a state where they can trade competitively at no cost. To keep the model simple, we indeed disregard physical trading costs.

We investigate a continuous two-state Markov-regime model, meant to formalize the idea that investors can either trade in the intermediated market or wait until a counterpart in the decentralized market arrives and trade with them. In the first regime or state s_1 the investor can trade only with the specialist and undergoes transaction costs $1-s$,$1/q-1$. In the second regime or state s_2 he can match his trade with other investors and transact without costs. The switching among the two states $X_t = s_1, s_2$ is governed by a Markov transition matrix Q where the entries $Q_{i,j}$, $i, j \in 1, 2$ are defined as

$$Q_{i,j} = \begin{cases} \liminf_{h\to 0} \frac{1-P(X(t+h)=s_i|X(t)=s_j}{h}) & i = j \\ \liminf_{h\to 0} \frac{P(Xt+h=s_i|X(t)=s_j)}{h} & i \neq j \end{cases}$$

We specify Q so that $-Q_{i,i} = Q_{i,j}$, $i \neq j$:

$$Q = \begin{pmatrix} -\lambda_1 & \lambda_1 \\ \lambda_2 & -\lambda_2 \end{pmatrix}$$

$$\lambda_1 \, \lambda_2 > 0.$$

This means that conditional on being in state s_1, s_2 at time t, the regime process is Poisson with instantaneous switching intensity λ_1, λ_2. The transition probability from one state to the other is

$$P(X(t+h) = j|X(t) = i) = \delta_{ij} + Q_{ij}h + o(h),$$

where δ_{ij} is the Kronecker delta. The stationary distribution, i.e. the long run proportion the process spends in states s_1 and s_2 for $t \to \infty$ is $\pi = (\lambda_2, \lambda_1)$.

7.1.1 The investor problem

The maximization problem for the investor becomes a system in the two value functions J^{i_1}, J^{i_2} - which apply respectively when starting from state s_1 and s_2 - which can be written and solved as in Dimitrakas (2008):

$$\begin{cases} \underset{z}{max} \left\{ rxJ_x^{i_1} + \mu y J_y^{i_1} + \frac{\sigma^2}{2} y^2 J_{yy}^{i_1} - (\beta + \lambda_1) J^{i_1} + \lambda_1 J^{i_2} \right\} = 0 \\ \underset{A(L_t,U_t)}{max} \left\{ rxJ_x^{i_2} + \mu y J_y^{i_2} + \frac{\sigma^2}{2} y^2 J_{yy}^{i_2} - (\beta + \lambda_2) J^{i_2} + \lambda_2 J^{i_1} \right\} = 0 \end{cases} \tag{24}$$

Assume that[22] $J^{i_1} = (x+y)^\gamma$ and $J^{i_2} = x^\gamma \boldsymbol{K}(y/x)$ by homotheticity; transform the variables as follows: $\theta = y/x,\ y = \theta/(1+\theta)W,\ x = 1/(1+\theta)W$ and substitute. From (24) we get the value of $\boldsymbol{K}$ as the solution of the corresponding homogeneous equation plus a particular solution [23]K_p of the complete equation:

$$\boldsymbol{K}(\theta) = A\theta^{\rho_1} + B\theta^{\rho_2} + K_p(\theta). \tag{25}$$

Here A, B are two constants, ρ_1, ρ_2 are the roots of the characteristic polynomial of the second degree equation:

$$(\delta - \lambda_1) + (\mu - r)\,\rho + \frac{\sigma^2}{2}\rho(\rho - 1) = 0.$$

and $\delta = r\gamma - \beta$ as in the previous sections.

In the no-cost state the investor should keep his portfolio at the optimum ratio θ^* dictated by the Merton's solution. Substituting K in the first equation and writing down the first order condition for the max with respect to θ, we get the following two equations:

$$\left[(1-\gamma)\left((\mu - r)\,\tfrac{\theta^*}{1+\theta^*} - \gamma\tfrac{\sigma^2}{2}\left(\tfrac{\theta^*}{1+\theta^*}\right)^2\right) - \delta - \lambda_1\right] C = -\lambda_1 \tfrac{\boldsymbol{K}(\theta^*)}{(1+\theta^*)^{1-\gamma}} \tag{26}$$

$$\left[\mu - r - 1 - \gamma\sigma^2 \tfrac{\theta^*}{1+\theta^*}\right] C = \lambda_1 \tfrac{\boldsymbol{K}(\theta^*)}{(1+\theta^*)^{-\gamma}} + \tfrac{\lambda^1}{\gamma-1}\tfrac{\boldsymbol{K}'(\theta^*)}{(1+\theta^*)^{-\gamma-1}} \tag{27}$$

By value-matching and smooth-pasting, the boundary conditions at lower and upper levels $\theta = l, u$ of the no-transaction zone are:

$$\begin{aligned}
\boldsymbol{K}'(l) &= \frac{\gamma}{q+l}\boldsymbol{K}(l) && (28)\\
\boldsymbol{K}''(l) &= \frac{\gamma - 1}{q+l}\boldsymbol{K}'(l) && (29)\\
\boldsymbol{K}'(u) &= \frac{\gamma}{1/s+u}\boldsymbol{K}(u) && (30)\\
\boldsymbol{K}''(u) &= \frac{\gamma - 1}{1/s+u}\boldsymbol{K}'(u) && (31)
\end{aligned}$$

Hence, the investor's problem is solved when eq. (26)-(31) are satisfied with $\boldsymbol{K}$ given by (25).

7.1.2 The specialist problem

Also for the specialist there are two different value functions, depending on the state he starts from. Since the specialist does not transact in state s_1, there is

[22] Notice that J^{i_1} is homogeneous of degree γ and at the optimum $J^{i_1}_x = J^{i_1}_y$

[23] $K_p(\theta) = \frac{2\lambda 2C}{\sigma^2(\rho_1 - \rho_2)}\left(z^{\rho_2}\int_{\theta_M}^{\theta}\frac{(1+t)^{1-\gamma}}{t^{\rho_2+1}}\mathrm{d}t - \theta^{\rho_1}\int_{\theta_M}^{\theta}\frac{(1+t)^{1-\gamma}}{t^{\rho_1+1}}\mathrm{d}t\right).\theta_M$ is the initial condition set to the Merton's solution.

no optimization in state $s_1$and he can optimize only in state s_2. The two value functions J^{s_1},J^{s_2} are defined as:

$$\begin{aligned} J^{s_1}(x,y,t;T) &\doteq \lim_{T\to\infty} \mathbb{E}\left[e^{-\beta'(T-t)}W(T)^{\gamma}/\gamma\right] && (32)\\ J^{s_2}(x,y,t;T) &\doteq \lim_{T\to\infty} \sup \mathbb{E}\left[e^{-\beta'(T-t)}W(T)^{\gamma}/\gamma\right] && (33) \end{aligned}$$

The system of equations which characterize these value functions can be written as:

$$\begin{cases} J^{s_1}_x rx + J^{s_1}_y \mu y + J^{s_1}_{yy}\sigma^2 y^2/2 - (\beta'+\lambda_1)J^{s_1} + \lambda_1 J^{s_2} & =0 \\ \underset{s,q}{max}\left\{J^{s_2}_x rx + J^{s_2}_y \mu y + J^{s_2}_{yy}\sigma^2 y^2/2 - (\beta'+\lambda_2)J^{s_2} + \lambda_2 J^{s_1}\right\} & =0 \end{cases} \qquad (34)$$

and is subject to the same BCs of the single-state case, i.e. (20) and (21). The first equation is the discounted Feynman-Kac equation - since no optimization occurs in state s_1 - while the second one is a Hamilton-Jacobi-Bellman equation valid under optimality. The costs, s, q are determined through the specialist optimization with (20) and (21). The optimization conditions for the specialist can no longer be solved explicitly, and the derivatives of l, u with respect to the costs s, q must be computed numerically. To solve the system, it is still possible to write $J^{s_{1,2}} = x^{-\gamma}I^{s_{1,2}}(\ \theta_s)$ and transform the system itself into two differential equations in $I^{s_{1,2}}(\theta_s)$.

$$\begin{cases} (\delta'-\lambda_1)I^{s_1} + (\mu-r)\theta_s I^{s_1\prime} + \frac{\sigma^2}{2}\theta_s^2 I^{s_1\prime\prime} + \lambda_1 J^{s_2} = 0 \\ (\delta'-\lambda_2)I^{s_2} + (\mu-r)\theta_s I^{s_2\prime} + \frac{\sigma^2}{2}\theta_s^2 I^{s_2\prime\prime} + \lambda_2 J^{s_1} = 0. \end{cases} \qquad (35)$$

The system (35) in I^{s_1} and I^{s_2} can be solved obtaining I^{s_1} as a function of I^{s_2} from the second equation and substituting it in the first equation. The first equation becomes a differential equation of 4^{th} order in I^{s_2}. Solutions are in the form:

$$I^{s_2} = c_1\theta_s^{x_1} + c_2\theta_s^{x_2} + c_3\theta_s^{x_3} + c_4\theta_s^{x_4} \qquad (36)$$

where $c_i, i = 1, ..4$ are constant, $x_{1,2} = \varsigma \pm \sqrt{\nu 1}$, $x_{3,4} = \varsigma \pm \sqrt{\nu 2}, \varsigma = (r - \alpha - 1/2\sigma^2)/(2\sigma^2)$

$$\begin{aligned} \nu 1 &= \frac{\sqrt{(-2r-2\alpha-\sigma^2)^2 - 8\delta'\sigma^2}}{2\sigma^2} && (37)\\ \nu 2 &= \frac{\sqrt{(-2r-2\alpha-\sigma^2)^2 - 4(-2\lambda_1 - 2\lambda_2 + 2\delta')\sigma^2}}{2\sigma^2} && (38) \end{aligned}$$

Hence, the specialist' s problem is solved by (36), provided that the constant δ' solves (18) and s, q satisfy (20), (21). There are three relevant intervals for δ'. In one of them the solutions for $x_{1,2,3,4}$, are all reals; in a second there are two reals and two imaginary solutions. In the third the solutions are all imaginary. We report here the form of the value function in the last interval, since this is the case which occurs in our numerical experiments below. With all imaginary solutions for x_i, the value function I^{s_2} can be written as:

$$I^{s_2} = \theta^{\varsigma}\left(A'si(\nu 1\log(\theta_s)) + B'co(\nu 1\log(\theta_s)) + C'si(\nu 2\log(\theta_s)) + D'co(\nu 2\log(\theta_s))\right)$$

We set $A' = B' = C' = D' = 1$ to avoid extra degrees of freedom. I^{s_1} can be found by substituting I^{s_2} in the second equation of (36).

7.1.3 Equilibrium

In order to find a solution, we need to solve for the nine unknowns δ, δ', s, q, l, u, A, B, m the system of 9 equations (18,20,21,26,27,28,29,30,31). This is the specialist-investor equilibrium when investors have the outside option to wait and trade without costs.

7.2 Numerical results

The numerical method to solve the system is illustrated in Appendix C.

7.2.1 Base-case

We explore solutions for the parametric base-case above, namely $\gamma = 4$,$\alpha - r = 5\%$,$\sigma^2 = 4\%$, $\gamma' = 3.85$. We take as switching parameters $\lambda_1 = 0.8$, $\lambda_2 = 0.2$, which implies a stationary distribution $\pi = (0.2, 0.8)$. The parameters λ_1, λ_2 are therefore chosen so that the system spends 1/4 of time in the costless state when $t \to \infty$. We get

$$\left(\delta, \delta', s, q\right) = (0.0355, 1.1352, 89.17\%, 90\%),$$

with barriers equal to

$$l = 0.1448 < \theta^* < u = 1.2337.$$

In order to illustrate the differences with respect to the single-state, or no-outside-option case, let us recall that there the main endogenous quantities were

$$\left(\delta, \delta', s, q\right) = (0.023428, 0.023687, 97.53\%, 68.41\%),$$

with barriers equal to

$$l = 0.301825 < \theta^* < u = 0.480013.$$

As expected from the elimination of the "monopolistic" position of the specialist, the welfare of the specialist decreases with respect to the single-state case. His δ' goes from 2.3% to 113%, which means that the rate of growth of his derived utility is drastically reduced. The welfare of the investor does not improve: his δ moves in the same direction as the specialist's one. However, the change is much smaller, from 2.3% to 3.5%. This happens because there are states in which he does not pay the price of immediacy, but he looses the intermediary's service of immediacy too. The slight change in his welfare signals that the two effects are almost comparable.

Since s is lower and q is higher in the new equilibrium, the overall transaction costs go down. Transaction costs at the upper barrier go from $1 - s = 2.5\%$ to 10.83% , while those at the lower barrier go from $1/q - 1 = 46.2\%$ to 11.11% . So, overall transaction costs $1/q - s$ go from 48.7% to 21.29%. Not only overall costs go down, but they are more symmetric at the two barriers, since upper costs increase little, lower costs dramatically decrease.

As for the intervention barriers, they widen, as intuition would suggest. The investor is more tolerant with respect to current discrepancies from the Merton's ratio, since he has a chance of being able to transact without costs in the future. The boundaries still contain the Merton's ratio, where the investor optimally sets his portfolio when no costs exist.

We can compare with the single-state base-case also the other relevant feature of the equilibrium, namely trade frequency. The first time to trade requires a more sophisticated computation than in the single-state case, since trade can occur in both states. Aa a preliminary result useful just for the sake of comparison, we compute the first time-to-trade conditionally on remaining in the cost case (as if there were no other state, which is evidently a very rough upper bound). This time, which was 0.5 before, raises up to almost 23 years.

7.2.2 Sensitivity analysis

Let us now vary the risk aversion of the specialist, in order to see how much this affects this "non-monopolistic" equilibrium and its comparison with the "monopolistic" case. In Table 3 below we report the equilibria parameters for the two cases - labeled respectively as "non-mon" and "mon" - when the specialist's risk aversion gets further from the investor's one. For the monpolistic case we report the figures down to $1 - \gamma' = 3.6$, while for the non-monopolistic case we go even further, to $1 - \gamma' = 3$, where the monopolistic equilibrium would give very high transaction costs.

Table 3

model	$1-\gamma'$	s	q	l	u	t^*	δ'	δ
mon	3.85	0.975	0.684	0.3018	0.4800	0.508	0.0236	0.0234
mon	3.6	0.841	0.560	0.2435	0.5647	2.973	0.0249	0.0234
non-mon	3.85	0.8917	0.9000	0.1448	1.2337	4.554	1.1352	0.0355
non-mon	3.6	0.8920	0.9000	0.1447	1.2315	4.544	1.1359	0.0355
non-mon	3	0.8328	0.9002	0.1449	1.8172	4.785	1.1400	0.0354

t^* for the non-monopolistic case is considered the time of first intervention after a switching to state s_2. It is computed as the minimum time between the transaction time to state s_1 (costless intervention) and the time of the crossing of one the barriers (costly intervention). The table confirms that, independently of risk aversion, the investor's option to wait and trade at no costs forces the specialist to accept lower transaction costs. The bid-ask prices behavior as a function of the difference between the agents' risk aversions $\gamma' - \gamma$ is presented in figure 4 below. The reader can appreciate the different level and symmetry of costs with respect to the monopolistic equilibrium

[insert here figure 4]

The non transaction cone $l \leq \theta \leq u$ is larger and the expected first trade time is higher in the non-monopolistic case. The effect on costs is sizable and welfare of the intermediary is drastically reduced. However, investors adjusts the barriers - and the time to trade - so that the effect on their welfare is almost none.

8 Summary and conclusions

We characterized equilibrium bid-ask spreads and infrequent trade in symmetric-information, intermediated markets. We actually specified two cases: either investors are obliged to trade with the specialist and incur into transaction costs, comprehensive of physical costs, or they can wait until another trader - with whom they can trade at no cost - arrives. In each economy, we provided the optimality conditions for market participants. These conditions determine the equilibrium bid and ask spreads, as well as the value functions of the agents and intervention barriers - or trade - of the investor.

We studied first the equilibrium in which trade occurs with the specialist only. In equilibrium, trade is the local time of the Brownian motion at appropriate levels, namely the trading barriers of the investor. We proved numerically that the equilibrium exists, at least for some combinations of risk aversion of its participants, and that its bid-ask spreads and trade frequency increase with the difference in risk aversions of the specialist and investor. The analysis was conducted both in the absence and in the presence of external trading costs.We then extended the analysis to the case in which investors can also wait and transact without paying the costs.

Our major contribution consists in endogenizing spreads and infrequent trade. With no outside option, effects on spreads are one order of magnitude bigger than their causes. Both with and without trading costs, intermediation imposes a price for immediacy which is very high in comparison to its motivation, i.e. difference in risk aversion, and very sensitive to changes in risk attitudes. Also the departure of the barriers from the optimal portfolio mix in the absence of costs is one order of magnitude bigger than the difference in risk aversion between market participants. Trade is infrequent, less than assumed by partial equilibrium models, but so as to wash away the continuous readjustments we often assume in continuous-time Finance. A small heterogeneity in risk aversion, together with a monopolistic position of the specialist, is able to produce high spreads and trade frequency of the order of months. The result is encouraging, given the low level of risk-aversion heterogeneity observed in recent empirical work. It may also be considered too strong, since our spreads are high with respect to "observed" levels. In order to address this issue, we studied also the equilibrium in which investors have the outside-option to wait and trade competitively. As expected, this option reduces the magnitude of the bid-ask spreads, without wiping infrequent trade out. The effect on costs is strong and welfare of the intermediary is drastically reduced. Investors however adjusts the barriers so that the effect on their welfare - which in principle could be negatively affected too, since either they trade through the intermediary and pay his rent, or must wait until a counterpart comes - is almost null.

Starting from exogenous costs - instead of endogenous spreads - Lo *et al.* had already noticed the strong effect of costly trading on prices, and made it a sign of distinction of their theoretical contribution with respect to the previous literature. Since they were not working in an intermediated market, they explained the strong impact of trading costs on prices via high-frequency trading needs and fixed costs. Maintaining the hypothesis of highly frequent trading-needs, we explained first-order effects on prices and trade through the existence of an intermediated market with its price for immediacy.

9 Appendix A

The three steps for solving the optimization problem of the investor are as follows. First, we recognize that a candidate solution for the value function is either

$$I(\theta) = \theta^{-m}\left[Asi(\nu\ln(\theta)) + Bco(\nu\ln(\theta))\right] \tag{39}$$

where $A, B \epsilon \mathbb{R}$, si and co are the trigonometric sine and cosine, or

$$I(\theta) = \mathcal{A}\theta^{x_1} + \mathcal{B}\theta^{x_2} \tag{40}$$

where $\mathcal{A}, \mathcal{B} \epsilon \mathbb{R}, x_{1,2} = m \pm \nu$. The type of solution depends on whether, having defined

$$\delta_{c.} \doteq \frac{\left(\alpha - r - \sigma^2/2\right)^2}{2\sigma^2}, \tag{41}$$

we have $\delta > (<)\delta_c$. Indeed, the algebraic equation corresponding to (6), which provides the roots $x_{1,2}$, i.e.

$$\sigma^2 x^2/2 + \left(\alpha - r - \sigma^2/2\right)x + \delta = 0 \tag{42}$$

has imaginary solutions in the first case, real in the second.

Second, we substitute both the first and second order BCs into the ODE, so as to obtain a second degree equation for the optimal barriers l and u, through their transforms ε_l and ε_u. These are respectively the smaller and the bigger root of the following equation:

$$\delta + \gamma\left(\alpha - r\right)\varepsilon + \gamma\left(\gamma - 1\right)\sigma^2\varepsilon^2/2 = 0 \tag{43}$$

whose discriminant we denote as Δ :

$$\Delta \doteq \gamma^2(\alpha - r)^2 - 2\delta\gamma(\gamma - 1)\sigma^2 \tag{44}$$

Third, we make the determinant of the value-matching BCs, once written in terms of (39) or (40), and considered as equations in (A, B) or $(\mathcal{A}, \mathcal{B})$, equal to zero. This guarantees that the value function is non-null and stationary. The determinant is equated to zero by a proper choice of the artificial discount rate β, via δ. This means solving for δ the algebraic equation

$$a(l,q)b(u,s) - c(u,s)d(l,q) = 0 \tag{45}$$

whose entries are defined as in the text for the imaginary case. Analogous expressions hold for the real case. The solution requires substitution of the expressions for $l, u, \epsilon_l, \epsilon_u$ in terms of the parameters $\alpha - r, \sigma$ and δ itself.

10 Appendix B

In order to compute the derivatives in (19), we first use the definition of ε_l and ε_u, namely (7) and (8), we can determine explicitly the investor's barriers:

$$\begin{cases} l = \frac{N}{D-N} q, \\ u = \frac{N'}{D-N'} \frac{1}{s} \end{cases} \tag{46}$$

where

$$D = \gamma(\gamma - 1)\sigma^2, \tag{47}$$

$$N = -\gamma(\alpha - r) - \sqrt{\Delta}, \tag{48}$$

$$N' = N + 2\sqrt{\Delta}, \tag{49}$$

Based on them, dependence of l on q and u on s acts both directly and via the discount rate δ (which equates the determinant of the value-matching conditions to zero, and therefore depends on all the model's variables, including q and s)

$$\frac{\partial l}{\partial q} = \frac{1}{D-N} \left[N + q \frac{D^2}{(D-N)\sqrt{\Delta}} \frac{\partial \delta}{\partial q} \right] \tag{50}$$

$$\frac{\partial u}{\partial s} = \frac{1}{s(D-N')} \left[\frac{-N'}{s} - \frac{D^2}{(D-N')\sqrt{\Delta}} \frac{\partial \delta}{\partial s} \right] \tag{51}$$

Using the implicit function theorem to derive the discount rate sensitivities, $\frac{\partial \delta}{\partial q}, \frac{\partial \delta}{\partial s}$, one has:

$$\frac{\partial \delta}{\partial q} = -\frac{b(u,s)\frac{\partial a(l,q)}{\partial q} - c(u,s)\frac{\partial d(l,q)}{\partial q}}{\frac{\partial (ab-cd)}{\partial \delta}} \tag{52}$$

$$\frac{\partial \delta}{\partial s} = -\frac{a(l,q)\frac{\partial b(u,s)}{\partial s} - d(l,q)\frac{\partial c(u,s)}{\partial s}}{\frac{\partial (ab-cd)}{\partial \delta}}. \tag{53}$$

where the derivatives of a, b, c, d are easily obtained in closed form (separately for the imaginary and real case). Putting the two together we have

$$\frac{\partial l}{\partial q} = \frac{1}{D-N}\left[N - q\frac{D^2}{(D-N)\sqrt{\Delta}}\frac{b(u,s)\frac{\partial a(l,q)}{\partial q} - c(u,s)\frac{\partial d(l,q)}{\partial q}}{\frac{\partial(ab-cd)}{\partial \delta}}\right] \tag{54}$$

$$\frac{\partial u}{\partial s} = \frac{1}{s(D-N')}\left[\frac{-N'}{s} + \frac{D^2}{(D-N')\sqrt{\Delta}}\frac{a(l,q)\frac{\partial b(u,s)}{\partial s} - d(l,q)\frac{\partial c(u,s)}{\partial s}}{\frac{\partial(ab-cd)}{\partial \delta}}\right] \tag{55}$$

which need to be substituted in the "modified" smooth pasting conditions as well as in conditions (20) and (21). The latter enter into the equilibrium computation.

11 Appendix C

The system of 9 equations (18,20,21,26,27,28,29,30,31) results to be time consuming to be solved in one step, especially in computing the numeric derivatives in (20), (21). We noticed that the investor and the specialist problems are coupled through s, q, l, u and the derivatives $\partial l/\partial q$, $\partial u/\partial s$. In particular, given s, q it is possible to solve the investor's problem (26,27,28,29,30,31). We solve the system in three steps. At the first step we solve the investor's problem computing l, u, $\partial l/\partial q$, $\partial u/\partial s$ at every point on a grid of values of $0.75 < s, q < 1$ with step 0.001. Even with a high number of points on the grid to be computed, this method allows us to solve a smaller problem and starting from chosen initial parameters. To initialize the parameter estimation, we fitted a linear relationship between s and u and between q and l. This allowed us to use the Quasi-Newton local search method. Moreover it permitted to compute the numerical derivatives only across the grid. The solution of the investor's problem is an array of four values l, u, $\partial l/\partial q$, $\partial u/\partial s$ for every point on the grid. At the second step we fit four splines curves for each of these variables. At the third step we solve the investor's problem (18,20,21) utilizing the splines fitting in place of l, u, $\partial l/\partial q$, $\partial u/\partial s$. After finding a solution, we refine and check the values of the solution solving the initial system of 9 equation all together, using as initial guess the solution found at step three.

12 References

Amihud, Y. and H. Mendelson, 1986, Asset Pricing and the Bid-Ask Spread, *Journal of Financial Economics,* 17, pp. 223-49.

Buss, A., and B. Dumas, 2012, The Equilibrium Dynamics of Liquidity and Illiquid Asset Prices, *Working Paper.*

Constantinides, G., 1986, Capital Market Equilibrium with Transaction Costs, *Journal of Political Economy,* 94, pp. 842-862.

Dimitrakas, V., Portfolio selection with stochastic transaction costs, PhD dissertation, INSEAD, 2008.

Duffie, D., N. Gârleanu and L.H. Pedersen, 2005, Over-the-Counter Markets, *Econometrica*, 73, pp. 1815-1847.

Duffie, D., N. Gârleanu, and L.H. Pedersen, 2007, Valuation in Over-the-Counter Markets, *Review of Financial Studies*, 20, pp. 1865-1900.

Dumas, B. and E. Luciano, 1991, An exact solution to a dynamic portfolio choice problem with transaction costs, *Journal of Finance*, 46, pp. 577-595.

Gerhold, S., Guasoni, P., Muhle-Karbe, J., and W. Schachermayer, 2011, Transaction Costs, Trading Volume, and the Liquidity Premium, unpublished working paper.

He, H., and H. Leland, 1993, On Equilibrium Asset Price Processes, *The Review of Financial Studies,* 6, pp. 593-617.

Ho, T., and H. Stoll, 1981, Optimal Dealer Pricing under Transactions and Return Uncertainty, *Journal of Financial Economics*, 9, pp. 47-73.

Jang, B.G., Koo, H.K., Liu, H., and M. Loewenstein, 2007, Liquidity premia and transaction costs, *Journal of Finance,* 62, pp. 2329- 2366.

Liu, H., and M. Loewenstein, 2002, Optimal Portfolio Selection with Transactions Costs and Finite Horizons, *The Review of Financial Studies,* 15, pp. 805-835.

Lo, A., H. Mamaysky and J. Wang, 2004, Asset Pricing and Trading Volume under Fixed Transaction Costs, *Journal of Political Economy,* 112, pp. 1054-1090.

Madhavan, A., and S. Smidt, 1991, A Bayesian model of intraday specialist pricing, *Journal of Financial Economics,* 30, pp. 99-134.

O'Hara, M., 1997, *Market microstructure theory*, Basil Blackwell, Cambridge, MA, 2nd ed.

Pagano, M., and A. Roell, 1993, Auction markets, dealership markets and execution risk, ch 8 in *Financial market liberalization and the role of banks*, ed. by Hamaui, R., Conti, V., Cambridge University Press.

Stoll, H., 1978, The supply of broker services in Securities Markets, *Journal of Finance*, 33, 1133-1151

Vayanos, D., 1998, Transaction Costs and Asset Prices: A dynamic equilibrium model, *Review of Financial Studies*, 11, pp. 1-58.

Xiouros, c, and F. Zapatero, 2010, The Representative Agent of an Economy with External Habit Formation and Heterogeneous Risk Aversion,*The Review of Financial Studies,* 23, pp. 3017-3047.

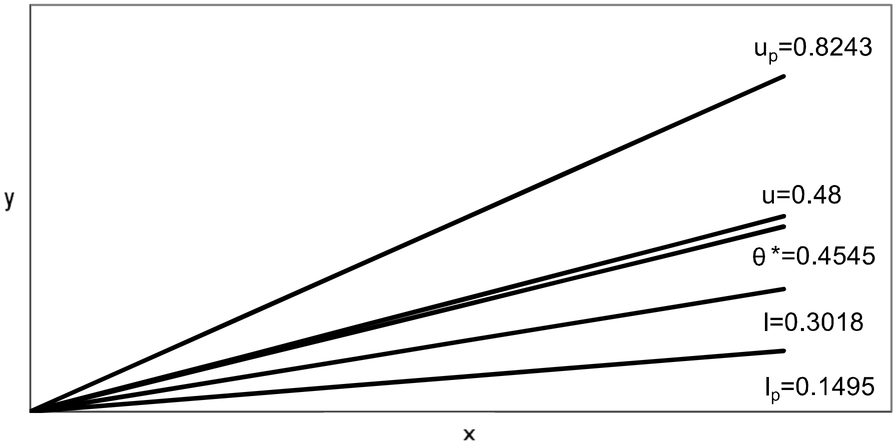

Figure 1: No-transaction cone in the single-agent (l_p, u_p) and monopolistic, general-equilibrium case (l, u). In both cases the optimal ratio of risky to riskless assets for the frinctionless market, θ^*, is included in the cone. The cone is in the plane of the asset values (x, y).

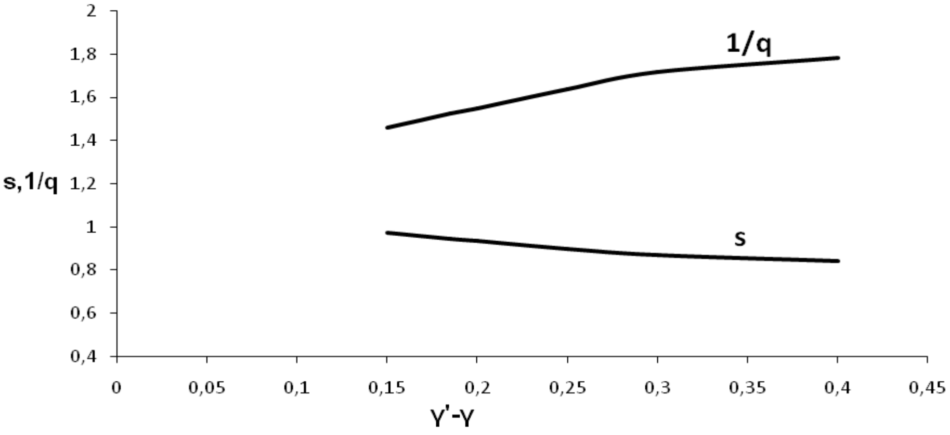

Figure 2: Bid price s and inverse of the ask price q as a function of the difference between the investor's and broker's risk aversion, $\gamma' - \gamma$ in the monopolistic general-equilibrium.

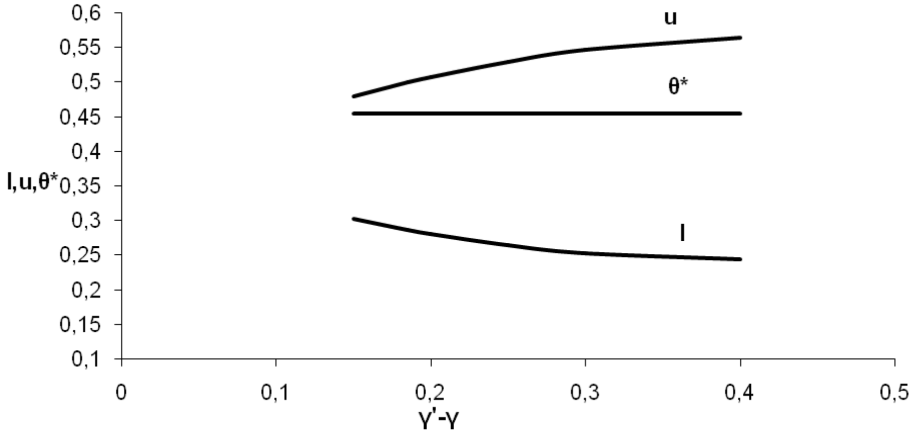

Figure 3: Cone of no-transactions at equilibrium as a function of the difference between the investor's and broker's risk aversion, $\gamma' - \gamma$. The cone is written in terms of the risky-to-riskless-asset ratio $\theta = y/x$, which stays between l and u. As in figure one, the cone contains the frictionless ratio θ^* in the monopolistic general-equilibrium.

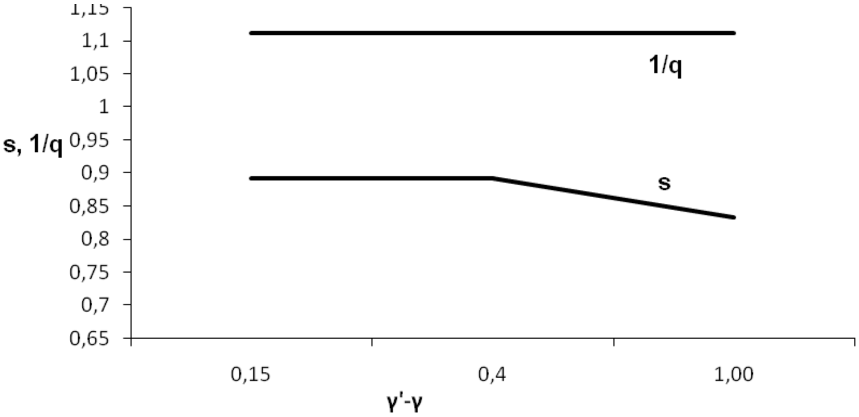

Figure 4: Bid price s and inverse of the ask price q as a function of the difference between the investor's and broker's risk aversion, $\gamma' - \gamma$ in the non-monopolistic general-equilibrium.