Journal of Fınancıal Economıcs 9 (1981) 47–73. North-Holland Publishıng Company

OPTIMAL DEALER PRICING UNDER TRANSACTIONS AND RETURN UNCERTAINTY*

Thomas HO

New York Universıty, New York, NY 10006, USA

Hans R. STOLL

Vanderbılt University, Nashville, TN 37240, USA

Receıved October 1979, final versıon receıved September 1980

The paper examınes the optimal behavior of a sıngle dealer who ıs faced wıth a stochastıc demand to trade (modeled by a continuous time Poısson jump process) and facıng return rısk on hıs stock and on the rest of his portfolio (modeled by dıffusion processes). Usıng stochastıc dynamic programmıng, we derıve the optimal bid and ask prıces that maxımıze the dealer's expected utility of termınal wealth as a functıon of the state ın which he finds himself. The relatıonshıp of the bıd and ask prıces to inventory of the dealer, ınstantaneous varıance of return, stochastıc arrıval of transactions and other varıables is examined.

1. Introduction

Dealers in securities markets stand ready to trade immediately fixed amounts of securities at stated bid or ask prices. In this paper the optimal bid and ask prices are determined such that the dealer's expected utility of terminal wealth is maximized. Our concern is primarily with the risk assumed by the dealer, which in this case arises not only from uncertainty about the return on his ınventory but also from the uncertainty about when future transactions will occur (which affects how long he must bear return uncertainty). We consider a single dealer trading a single stock and facing a downward sloping stochastic demand for the service of immediacy provided by the dealer.

This paper is related to an earlier paper by Stoll (1978a); indeed the economic setting of the problem is much the same. What is different is the nature of uncertainty (primarily the introduction of transactions uncertainty), explicit treatment of a multiperiod strategy for the dealer, and the introduction of the demand side. The paper should also be placed in the context of the growing literature on the micro-structure of securities markets.

*We wısh to thank Thomas Copeland, E. Phılıp Jones, Robert Schwartz and the referee for theır helpful comments

This literature is concerned with the difference between transactions prices and 'true' underlying prices as determined by the available information set. One approach to this issue is to propose price or return generating processes that are consistent with our understanding of market structure but do not necessarily derive from any principles of individual economic behavior. The end product of such research is the relationship between the observed distribution of returns and the 'true' underlying distribution. The work of Clark (1973), Blattberg and Gonedes (1974), Westerfield (1977), Oldfield, Rogalski and Jarrow (1977) Schwartz and Whitcomb (1977), and Goldman and Beja (1979) is in this spirit. The latter paper is most closely related to our paper since it is concerned with the impact of the specialist on the process by which transactions prices adjust to the evolving 'true' price of the stock. However, they do not build an explicit micro-economic model of the dealer.

An alternative approach is to specify the demand for and supply of the services of intermediaries — i.e., dealers — whose compensation is reflected in a transaction price different from the 'true' price. For an extensive review of the literature, see Cohen, Maier, Schwartz and Whitcomb [CMSW] (1979). On the supply side, work began with the paper by Demsetz (1968), and additional theoretical contributions were made by Tinic (1972) and Stoll (1978a). There are also a host of empirical papers on bid–ask spreads that have arisen out of this or closely related approaches. In addition to Demsetz and Tinic, see Tinic and West (1972), Benston and Hagerman (1974), Stoll (1978b), and Smidt (1979), for example. Papers by Copeland (1976), Epps (1976), CMSW (1978), and Goldman and Sosin (1979) can be viewed as the basis for the demand for dealer services because they specify why investors trade and how trading affects prices. However, the role of intermediaries is not explicitly considered.

In an excellent paper, Garman (1976) does consider the role of the intermediation process, but he is more concerned with the equilibrium price of *securities* rather than the equilibrium price of the *services* of intermediaries. We take the dealer's opinion of the 'true' price of the stock to be exogenously determined by his information set and ask how the dealer prices relative to his 'true' price whereas Garman is concerned with the necessary behavior of transactions prices to meet an exogenous inventory objective of the dealer. We adopt Garman's concept of stochastic supply of and demand for securities but view both the demand to sell securities and the demand to buy securities as demands for dealer services. Unlike Garman and recent papers by Amihud and Mendelson (1979) and Mildenstein and Schleef (1979), based on Garman, we do not require our dealer to maximize expected profits. Indeed our key concern is with the risk the dealer faces, and how this affects his willingness to supply dealer services. In addition we introduce stochastic processes for the 'true' return on the dealer's stock and the other components of his portfolio. In the face of stochastic transactions and stochastic returns, the dealer maximizes expected utility of terminal wealth by adjusting bid and ask prices through time. We present a dynamic programming solution to this more general problem. Mildenstein and Schleef also employ dynamic programming but they assume a risk premium for the dealer rather than deriving the premium in their analysis.

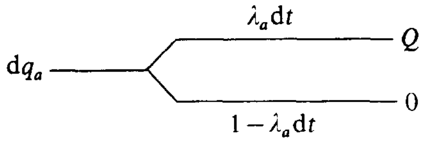
This paper is limited to the behavior of a single dealer making a market in a single stock. The generalization to many stocks is not difficult but adds needless complexity and is not considered in this paper. The determination of market bid and ask prices when there are several dealers (or, for that matter, individuals placing limit orders) is also considered elsewhere.[1]

The next section of this paper discusses the assumptions underlying our model. The dynamic programming problem is then stated and a solution presented. The model is used to analyze the dealer's optimal bid–ask spread and price adjustment strategy under symmetric as well as asymmetric demand conditions. A numerical illustration is also presented. The model is then used to analyze the decision to become a dealer in a particular stock. Finally the paper is summarized.

2. Assumptions

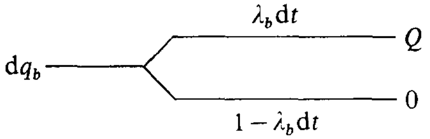
Transactions are assumed to evolve as a stationary continuous time stochastic jump process as in Garman (1976). A different process is allowed for purchases by the dealer and sales by the dealer. The process may be represented as follows:

Dealer sales (public purchases):



(1a)

Dealer purchases (public sales):



(1b)

where Q is the jump size (number of shares in a transaction), and λ_a, λ_b are the intensities of processes, representing the average number of public purchase or public sale transactions respectively per unit time. $\lambda_a \mathrm{d}t$ and $\lambda_b \mathrm{d}t$ can be interpreted as the probabilities of a dealer sale or a dealer purchase over the next instant.

[1] A simplified version of our approach and some results are in Ho and Stoll (1980).

The use of a Poisson jump process to model transactions is quite natural. In actual markets, dealers are obligated to trade a minimum quantity (i.e., 100 shares), and transactions occur at discrete time intervals.

The dealer determines a price of immediacy, b, should a public sale (dealer purchase) order arrive and a price, a, should a public purchase (dealer sale) order arrive. The dealer does not directly quote b, a. Rather he quotes bid and ask prices which are defined as

$$p_b = p - b, \tag{2a}$$

$$p_a = p + a, \tag{2b}$$

where p is the dealer's opinion of the true price of the stock at the time he sets the bid–ask quotation. The public's opinion of the true price is denoted by p^*, and the two need not be the same.

The dealer is assumed to trade only with the public in a passive way. That is, he sets prices and must wait until a transaction occurs. He cannot immediately dispose of inventory by trading with a dealer of last resort. The arrival rate of dealer purchases, λ_b, and dealer sales, λ_a, depends on the dealer's buying fee and selling fee respectively. Linear demand curves are shown in fig. 1. By raising b, the dealer can reduce the probability of further public sales and by lowering a, he raises the probability of public purchases. Of course public purchases and sales are also influenced by the difference between p and p^*. This difference is assumed to be parameterized in the functions λ_a and λ_b.

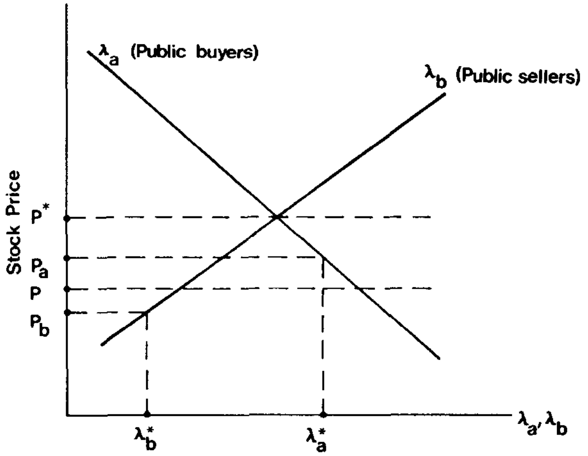


Fig. 1 Average public demand to buy from dealer, λ_a, and average public demand to sell to dealer, λ_b, when public's opinion of 'true' price (p^*) and dealer's opinion of 'true' price (p) differ Dealer sets bid price (p_b) and ask price (p_a) and regulates demand at λ_b^* and λ_a^*

In addition to uncertainty about the timing of subsequent transactions, the dealer faces uncertainty about the return on his existing portfolio. In the absence of a transaction, portfolio growth, dX, is represented by a stochastic differential equation of the following form:

$$\mathrm{d}X = r_X X\,\mathrm{d}t + X\,\mathrm{d}Z_X, \tag{3}$$

where r_X is the mean return per unit time; dZ_X is a (non-standardızed) Wiener process with mean zero and instantaneous variance, σ_X^2.

The dealer's wealth is divided into three components:

(1) *Cash*

Cash is accumulated when the dealer sells securities (short selling is permitted) and paid out when the dealer buys securities. Any balance in the cash account earns (or pays) the risk free rate of interest, r. The change in the value of the cash account, F, ıs

$$\mathrm{d}F = rF\,\mathrm{d}t - (p-b)\,\mathrm{d}q_b + (p+a)\,\mathrm{d}q_a. \tag{4}$$

Uncertainty in the cash account is due to uncertainty about transactions not to any uncertainty about the interest to be earned.

(2) *Inventory*

The dealer's inventory consists of shares of the one stock in which he makes a market. The change in the value of the inventory account, I, ıs

$$\mathrm{d}I = r_I I\,\mathrm{d}t + p\,\mathrm{d}q_b - p\,\mathrm{d}q_a + I\,\mathrm{d}Z_I, \tag{5}$$

where notation is as in (3) and (4). Shares enter inventory at p whereas the associated cash flow in the cash account is $(p-b)$ or $(p+a)$ per share. The difference reflects the dealer's profit. Uncertainty arises both from transactions uncertainty reflected in dq_a and dq_b and from uncertainty about the return on the stock which is reflected in the instantaneous variance σ_I^2, of the Wiener process, dZ_I. Over time, the dealer experiences discrete jumps in inventory due to transactions of Q shares at price p. There ıs an offsetting jump of slightly smaller size (smaller by the transaction cost, a or b) in the cash account. The fluctuations in value between jumps are due to return uncertainty represented by the last term in (5), and this return uncertainty is the underlying source of risk facing the dealer.

In the real world, fluctuations in return are reflected in the price of the stock (absent dividend payments). In our model this would require a complex accounting through time of the relation between return fluctuations and price fluctuations. However, allowing the price to fluctuate introduces no new element of uncertainty that is not already reflected in return uncertainty. Thus we adopt the convention that the true price, p, is kept constant. This implies that return uncertainty shows up as a continuous stock dividend that keeps the true price constant. It should be emphasized that this assumption of constant p is an accounting convention and that no sources of uncertainty are ignored. The change in shares outstanding could ultimately affect the demand to trade in the dealer's stock; but the effect would be diffused across all stocks, and given the short time horizon of the model, would be unlikely to be consequential to the dealer's decision on how to set his spread today.

(3) *Base wealth*

The dealer has base wealth Y. On the day he starts as a dealer $F_0=0$ and $I_0=0$, and base wealth is his wealth. The dealer is assumed to have one consumption point at time T, and base wealth is an optimally diversified portfolio with respect to that time horizon. Base wealth serves as collateral to back up the borrowing of shares or money that arises in the course of acting as a dealer. The change in base wealth is

$$\mathrm{d}Y=r_Y Y\mathrm{d}t+Y\mathrm{d}Z_Y, \tag{6}$$

where notation is as in (3). Unlike Garman (1976), we are not concerned with bankruptcy of the dealer. We assume the time horizon is sufficiently short and the collateral is sufficiently large to make bankruptcy an unimportant consideration in light of the Wiener process in which sudden jumps in return are not possible and in light of the dealer's ability to adjust prices. The effect of bankruptcy in a model with extended time horizon and price jumps deserves further investigation.

The objective of the dealer is to maximize the expected utility of his total wealth, $EU(W_T)$, at time T, his horizon, where

$$W_T=F_T+I_T+Y_T. \tag{7}$$

The world ends at $T$and the dealer is assumed to liquidate his inventory and base wealth at their market values without transactions costs.[2] While consumption occurs once in our model, the multiperiod aspects are reflected in the random shocks to the opportunity set arising from numerous transactions and price changes prior to the horizon date. We seek the optimal strategy for choosing the values of a and b that maximize the dealer's preference function. This involves weighing the costs and benefits of

[2] We could impose an arbitary liquidation fee which would affect our mathematical results but would not add greatly to our understanding of the risks and appropriate dealer strategy in the periods prior to the consumption point

being a dealer. The costs arise from the non-optimal portfolio the dealer stands ready to accept. Not only is his degree of diversification and level of risk not optimal but the presence of transactions uncertainty makes uncertain when and how much his portfolio will be unbalanced. The benefits arise from the fee he is able to charge for his service of immediate trading. Because the model assumes a single dealer in the stock, that dealer has the ability to earn monopoly profits.

The optimal strategy is complicated by the multiperiod framework which permits the dealer to adjust a and b as he moves through time — usually in response to inventory changes. Although stochastic, these inventory changes are in turn influenced by the bid and ask prices. This is known as a closed loop control problem because the optimal values of a and b depend on the observed state variables (F, I, Y) as well as on time, t. In other words the optimal dealer *strategy* we seek is a *function* that specified the choice of a and b for any position (described by t, F, I, Y) in which the dealer finds himself. The appropriate procedure for such a problem is dynamic programming.

Two points about the relationship of our analysis to the literature on portfolio theory and capital asset pricing should be emphasized. First, ours is a partial equilibrium analysis that is concerned only with the individual dealer's portfolio problem. No assumptions are made about equilibrium asset prices and no solutions are offered to equilibrium asset pricing in the presence of a dealer. Second, our analysis does not assume multiperiod consumption. In our model, as in standard one period portfolio models, the dealer costlessly liquidates all assets at the horizon date. The model differs in the multiperiod evolution of the opportunity set. Despite the single consumption horizon, time dependencies in the opportunity set, induced by the dealer's bid and ask pricing strategy, make it impossible to reduce our problem to the standard one period problem. We leave to future research the development of a more complete multiperiod portfolio analysis of the dealer and the integration of dealer behavior into the theory of equilibrium asset pricing.

3. Statement of dynamic programming problem

For any state of the world which is described by time and the values of the state variables — dealer's cash, inventory and base wealth — the dealer chooses a and b to maximize the expected utility of his terminal wealth. Define the optimal performance function $J(\quad)$ as

$$J(t,F,I,Y)=\max_{a,b}\left[EU(W_T)\,|\,t,F,I,Y\right]. \tag{8}$$

The function $J(\ \ )$ is the solution to the maximization problem (8) for an optimal bid and ask strategy from t_0 to T. It can be viewed as a derived utility function which describes the result of optimal dealer behavior conditional on any observed t, F, I, Y. Since there is no intermediate consumption prior to T, the fundamental recurrence relation implied by the principle of optimality of dynamic programming is simply that

$$\max_{a,b} \mathrm{d}J(t,F,I,Y)=0, \tag{9a}$$

and

$$J(T,F,I,Y)=U(W_T). \tag{9b}$$

In other words J must meet the condition that the maximized increments to J are always zero; for if they were not, one could increase derived utility by an alternative bid–ask strategy. Also, at time T, J must give the same level of utility as the elementary utility function, U.

Writing out the partial differential equation implied by (9a) — Bellman's equation — gives

$$\begin{aligned}\max_{a,b}(\mathrm{d}J/\mathrm{d}t)=J_t+LJ&\\ &+\max_{a,b}\{\lambda_a[J(F+pQ+aQ,I-pQ,Y)-J(F,I,Y)]\\ &+\lambda_b[J(F-pQ+bQ,I+pQ,Y)-J(F,I,Y)]\}=0,\end{aligned} \tag{10}$$

where J_t is the partial derivative of J with respect to time and L is defined as the operator that takes the differential with respect to time of the mean returns and Wiener risk components of the function J using Ito's Lemma. Thus[3]

$$LJ=J_F rF+J_I r_I I+J_Y r_Y Y+\tfrac{1}{2}J_{II}\sigma_I^2 I^2+\tfrac{1}{2}J_{YY}\sigma_Y^2 Y^2+J_{IY}\sigma_{IY}IY, \tag{11}$$

where σ_I^2 and σ_Y^2 are the instantaneous variance of $\mathrm{d}Z_I$ and of $\mathrm{d}Z_Y$ respectively, and σ_{IY} is the covariance between $\mathrm{d}Z_I$ and $\mathrm{d}Z_Y$, and where subscripts on a function indicate the variable with respect to which a partial derivative is taken. As is evident in (4), (5) and (6), the mean returns and Wiener risks do not depend directly on a and b, and therefore LJ does not involve maximization over a and b. Thus J_t+LJ simply represents the total

[3] A source on the dynamic programming solution to problems in stochastic control is Davis (1977). Davis also uses the operator L in the sense we do To illustrate the operator L, consider a simplified problem. Suppose $y=f(t,x)$ where $\mathrm{d}x$ is given by (3). Using Ito's Lemma and taking expectations gives $E\,\mathrm{d}y=f_t\,\mathrm{d}t+f_X E\,\mathrm{d}x+\frac{1}{2}f_{XX}E(\mathrm{d}X)^2=f_t\,\mathrm{d}t+f_X E[r_X X\,\mathrm{d}t+X\,\mathrm{d}Z_X]+\frac{1}{2}f_{XX}X^2\sigma_X^2\,\mathrm{d}t$. Noting that $E\,\mathrm{d}y=\mathrm{d}Ey$ [see Davis (1977)] and that $E(X\,\mathrm{d}Z_X)=0$, it follows that $\mathrm{d}Ey/\mathrm{d}t=f_t+Lf$ where $Lf=f_X r_X X+\frac{1}{2}f_{XX}\sigma_X^2 X^2$, which is analogous in form to (11).

time derivative of derived utility when there are no transactions. (Of course a and b indirectly affect Y, I and thereby influence the dealer's ultimate return and risk.)

More important to the dealer and to our problem are the last two lines of (10) which represent the maximized increments to $J(\ )$ resulting from transactions the value of which are pQ and which occur with probability $\lambda_a \mathrm{d}t$ in the case of dealer sales or $\lambda_b \mathrm{d}t$ in the case of the dealer purchases. The dealer also adds his fee, aQ or bQ, to his cash account. These increments do depend on a and b. The independence of the processes, $\mathrm{d}q_a$ and $\mathrm{d}q_b$, allows us to write the effect as the sum of the two increments. The time argument in $J(\ )$ has been suppressed.

It is convenient to restate the problem in terms of time remaining to the horizon date T. Let τ denote this amount of time, and replace t in (8) by τ. In other words we now think of J as a derived utility function that depends on the time remaining to the horizon date. Thus when $\tau=0$,

$$J(0,F,I,Y)=U(W). \tag{12}$$

The derived utility at the horizon data is just the utility of wealth at that point. The transformation of the time variable also implies that

$$J_t=-J_\tau. \tag{13}$$

Next consider the last two lines of (10), and since p is arbitrary let $p=1$. Because aQ and bQ are small relative to Q, we make the following first-order approximations:

$$J(F+Q+aQ,I-Q,Y)=J(F+Q,I-Q,Y)$$
$$+J_F(F+Q,I-Q,Y)aQ, \tag{14a}$$
$$J(F-Q+bQ,I+Q,Y)=J(F-Q,I+Q,Y)$$
$$+J_F(F-Q,I+Q,Y)bQ. \tag{14b}$$

Note that there is no approximation with respect to the effect of a transaction on J; the approximation is only with respect to the fee collected on a transaction.

We assume the same linear demand relation for dealer sales and purchases:

$$\lambda_a=\lambda(a)=\alpha-\beta a, \tag{15a}$$

$$\lambda_b=\lambda(b)=\alpha-\beta b. \tag{15b}$$

This assumption of symmetric demand is made for expositional clarity and will be relaxed later. Now define the sell operator, S, as

$$SJ = S[J(F, I, Y)] = J(F+Q, I-Q, Y), \tag{16a}$$

and define the buy operator, B, as

$$BJ = B[J(F, I, Y)] = J(F-Q, I+Q, Y). \tag{16b}$$

The sell operator acting on J describes the dealer's derived utility after a sale by the dealer but before including the selling fee. Utility will decrease if the sale drives the dealer farther away from his desired portfolio by increasing a short inventory position. Utility will increase if the sale reduces a long inventory position. A corresponding result holds on the buy side.

Substituting (13), (14) and (15) into (10) and using (16) allows us to restate (10) as follows:

$$\begin{aligned} J_t = LJ + \max_{a,b} \{ & [\lambda(a) a Q SJ_F - \lambda(a)(J - SJ)] \\ & + [\lambda(b) b Q BJ_F - \lambda(b)(J - BJ)] \}. \end{aligned} \tag{17}$$

This equation says that the change with respect to time remaining of derived utility depends on the return and risk of the dealer's current wealth (LJ), over which he has no direct control, and the maximized value of two terms that reflect the net contribution to the dealer's derived utility from dealer sales and purchases respectively, over which he has control. The essence of the dealer's problem is contained in these two terms.

Consider the last term of (17), which represents the net benefit from a purchase by the dealer. The expected gross benefit (in utile dimensions) from a dealer purchase is

$$\lambda(b) b Q BJ_F,$$

where $\lambda(b)bQ$ is the expected revenue from purchase transactions; and BJ_F is the marginal utility of cash after a purchase transaction, i.e., BJ_F converts expected revenues into 'utiles'. Since Q and BJ_F are given and since $\lambda(b)$ has been assumed to be linear, the expected gross benefit is a quadratic function of b and is drawn in fig. 2. The curve passes through the origin, is at a maximum when $b = \alpha/2\beta$ [from (15)], and is zero again at $b = \alpha/\beta$. Because of the assumed symmetry in demand conditions and on the temporary assumption that $BJ_F = SJ_F$ the curve is identical for dealer sale transactions. The assumption $BJ_F = SJ_F$, which implies that the marginal derived utility of cash after a sale equals the marginal derived utility of cash after a purchase, is made for expositional purposes only and is not used in the formal development of the model.

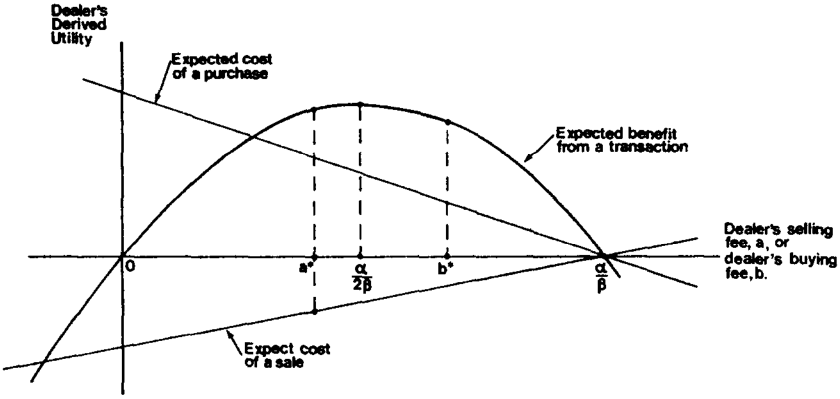


Fig. 2. Dealer's expected (gross) benefit function for a purchase or sale transaction $=\lambda(x)xQBJ_F$, with $x=a,b$, where a, b are the selling and buying fee respectively, $\lambda(\cdot)$ is the mean arrival rate of transaction, Q is the transaction size, and BJ_F is the marginal utility of the fee after the transaction. Dealer's expected (gross) cost function for a purchase, conditional on positive initial inventory$=(J-BJ)\lambda(b)$, where $(J-BJ)>0$ is the decrease on the dealer's derived utility resulting from a purchase. Dealer's expected (gross) cost function for a sale conditional on positive initial inventory$=(J-SJ)\lambda(a)$, where $(J-SJ)<0$ is the decrease in the dealer's derived utility resulting from a sale. Dealer maximizes expected, risk adjusted, (net) benefit by choosing optimal selling fee, a^*, and optimal buying fee, b^*. Risk neutral buying or selling fee is $\alpha/2\beta$.

What is typically different between an incremental purchase and an incremental sale is the gross 'cost' to the dealer (in utile dimensions). Unless the dealer begins with zero inventory, he prefers one type of transaction over the other. Suppose the dealer begins with positive inventory which makes him reluctant to purchase additional shares. The expected *decrease* in derived utility from a purchase,

$$\lambda(b)(J-BJ),$$

is positive because $(J-BJ)>0$. Since $\lambda(b)$ is linear and downward sloping this 'cost' function is also linear and downward sloping as shown in fig. 2. The larger b the lower is the probability of a dealer purchase and therefore the lower is the expected 'cost' to the dealer of a purchase. The optimum buying fee, b^*, is set to maximize the difference between the 'benefit' and 'cost' curve. Performing this maximization yields

$$b^*=\alpha/2\beta+(J-BJ)/2BJ_FQ. \tag{18a}$$

In other words, set b at the level which maximizes expected profits associated with the stochastic arrival of public sellers plus a risk premium term. The risk premium term ultimately depends on the dealer's fundamental attitude toward risk and the level of return and transactions uncertainty of the stock, all of which enter into J and BJ.[4] On the other hand, when inventory is positive, utility is *increased* by a sale transaction, i.e., $\lambda(a)(J-SJ)<0$ because $(J-SJ)<0$. The new position after a sale, SJ, would be superior to the old position, J. Thus the 'cost' function for a sale transaction is negative and upward sloping as shown in fig. 2. The larger is a, the lower is the probability of a dealer sale and therefore the less negative is the expected cost (the smaller is the expected increase in utility). The optimum selling fee, a^*, is a set to maximize the difference between the 'benefit' curve and the 'cost' curve for sales. Analogously to the case of a purchase this solution is

$$a^* = \alpha/2\beta + (J-SJ)/2SJ_F Q. \tag{18b}$$

Because the dealer has positive inventory, $b^* > a^*$ in fig. 2, as one would expect. When dealer inventory is zero, the two 'cost' functions coincide and are downward sloping, in which case $a^* = b^*$. As inventory changes, the 'cost' functions shift in opposite directions and a and b change in opposite directions. For given p, this implies that both the bid price (p_b) and ask price (p_a) fall in response to inventory increases and both rise in response to inventory decreases, as shown in fig. 3. We shall see later that the spread, the difference between p_a and p_b, is to an important degree invariant with respect to inventory changes.

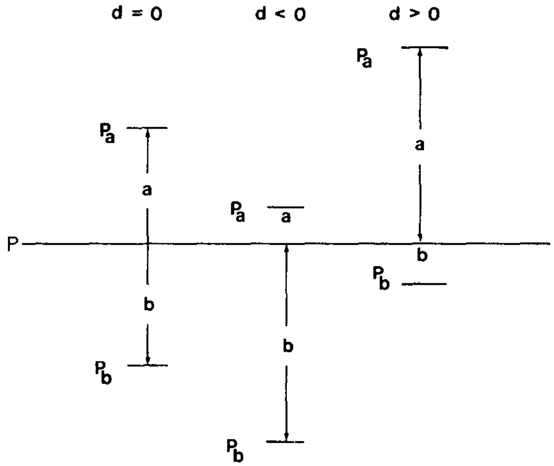


Fig 3 Adjustment of bid price (p_b) and ask price (p_a) relative to 'true' price, p, with constant spread, $s=a+b$, and changing price adjustment, $d=a-b$

[4] Garman's dealer would set $b^* = \alpha/2\beta$, and the analysis in that case is, for example, concerned with the time until bankruptcy. Amihud and Mendelson allow the dealer to adjust his price — in our model, a and b — to avoid failure. But they maintain the assumption of risk neutrality.

Until now our methodology has been quite standard. We have formulated the dynamic programming problem and stated the solution — (18) — in terms of marginal conditions on the derived utility function. However, we wish to go beyond the qualitative conclusions contained in our discussion of fig. 2 and eq. (18) and analyse the specific determinants of the dealer's bid–ask strategy in a multiperiod environment. This requires a solution for J.

We now proceed to a final statement of the dynamic programming problem by performing the maximization required in (17). This has in fact been done in (18a) and (18b). Substitution of (18) into (17) and using (15) yields

$$\begin{aligned} J_\tau = {} & LJ + [\alpha SJ_F Q - \beta(J - SJ)]^2/4\beta SJ_F Q \\ & + [\alpha BJ_F Q - \beta(J - BJ)]^2/4\beta BJ_F Q. \end{aligned} \tag{19}$$

The problem is to solve (19) for J subject to (18) and the terminal condition, (12). The solution for J can then be substituted into (18) to yield the overall optimal values of a and b as functions of the state variables, F, I, Y.

4. Solution

Unfortunately to our knowledge there is no closed form solution for equations of the form (19). However by using Taylor's series expansion we can, in principle, come arbitrarily close to the solution. Indeed this procedure has certain advantages since it allows one to see the nature of the solution as a function of time to the horizon date. Thus write $J(\ )$ as a Taylor's expansion around $\tau = 0$,

$$J(\tau, F, I, Y) = \phi^0(F, I, Y) + \phi^1(F, I, Y)\tau + \phi^2(F, I, Y)\tau^2 + \ldots, \tag{20}$$

where higher-order terms are assumed to be small enough to be neglected. It will be evident from the methodology that follows that higher-order terms could be incorporated. However, for the purpose of this paper the increased complexity that results is not offset by a corresponding increase in understanding of the dealer. The interpretation of (20) is that $J(\ )$ depends on how much time remains in the dealer's planning horizon. When $\tau = 0$, only the first term of (20) is relevant. When $\tau > 0$ but sufficiently small, only the second term is relevant, and so on.[5]

The solution to (20) is found by finding the different functions $\phi^0(\ )$, $\phi^1(\ )$ and $\phi^2(\ )$. The superscripts on the ϕ are notations associating each function with the degree of the derivative of J with respect to τ. Eq. (20) holds for all values of τ. For $\tau = 0$, J is given by (12). Therefore

$$\phi^0(F, I, Y) = U(W), \tag{21}$$

[5] As a practical matter the omitted highet order terms are not likely to be large for the short planning horizons over which most dealers operate

the utility of terminal wealth. In other words if the dealer is at maturity the optimal performance function is just his utility function.

For $\tau>0$, (19) must be satisfied. Substituting for J_τ, J and J_F in (19) from (20) and evaluating the resulting expression at $\tau=0$ gives the solution for $\phi^1(\ )$,[6]

$$\phi^1=LU(W)+2U'\Pi, \tag{22}$$

where

$$\Pi=\alpha^2Q/4\beta.$$

The operator, L, has been defined in (11). Its application in this case leads to

$$LU\,(W)=U'r_W W+U''\left(\tfrac{1}{2}\sigma_I^2I^2+\sigma_{IY}IY+\tfrac{1}{2}\sigma_Y^2Y^2\right),$$

where

$$r_W=r(F/W)+r_I(I/W)+r_Y(Y/W),$$

and $U'>0$ and $U''<0$ are the first and second derivatives of $U(W)$. In other words $LU(W)=L\phi^0$ represents the positive contribution of mean return to utility when there is only a short instant of time remaining plus the negative contribution of return variance. The term, Π, is the maximum expected profits from either sales or purchases (which are the same under symmetric demand assumed here).[7] The term $2U'\Pi$ thus represents the contribution to dealer welfare of the profits he expects to make from purchase plus sale transactions. Transactions uncertainty does not enter into ϕ^1.

[6] From (20),

$$(J_\tau\,|\,\tau=0)=\phi^1, \qquad (J\,|\,\tau=0)=\phi^0=U, \qquad (J_F\,|\,\tau=0)=\partial\phi^0/\partial F=U'.$$

Thus (19) evaluated at $\tau=0$ is

$$\phi^1=LU+[\alpha SU'Q-\beta(U-SU)]^2/4\beta SU'Q+[\alpha BU'Q-\beta(U-BU)]^2/4\beta BU'Q.$$

Because $U=SU=BU$ and $SU'=BU'=U'$ (since at the horizon date there is no risk and the utility of a dollar is the same whether held in inventory or cash),

$$\phi^1=LU+2U'(\alpha^2Q/4\beta), \quad \text{which is (22).}$$

[7] Using (15a) expected profit per sale transaction of size Q is

$$\lambda_aQ=\alpha Q-\beta Qa^2.$$

The value of a which maximizes this expression is $a=\alpha/2\beta$ Substituting this in the expected profit expression yields maximum expected profit equal to $\alpha^2Q/4\beta$ a value denoted by Π. Finally substituting the value, $\alpha/2\beta$, in the demand relation yields the optimal arrival rate at the price which maximizes expected profit, i.e., $\alpha/2$. Corresponding terms arise in the case of dealer purchases.

The function ϕ^2 is found in a similar way. Substitute for J_τ, J and J_F in (19) from (20), take the derivative with respect to time, and evaluate at $\tau = 0$. The solution uses the solutions for ϕ^0 and ϕ^1, and involves derivatives and moments of third and fourth order. To simplify the problem, we now assume the dealer's elementary utility function is quadratic so that we may restrict ourselves to a mean-variance world.[8] The solution for ϕ^2 is then[9]

$$\phi^2 = \tfrac{1}{2}L^2U(W) + U'r\Pi + 2U''\Pi^2 + 2U''\Pi r_W W + \tfrac{1}{2}U''\frac{\alpha}{2}\sigma_I^2 Q^2. \tag{23}$$

This expression is a negative adjustment to the dealer's derived utility for aspects of return uncertainty and the effect of transactions on portfolio return uncertainty. The first term of (23) reflects the compound effect of return and return risk resulting from twice applying the operator $L$to $U(W)$. The term is written out in the footnote and inspection will indicate that it is normally negative.[10] The second term is the increase in utility resulting from the interest earnings on the transaction profit. Because $U'' < 0$ the remaining terms are always negative. The third term, $2U''\Pi^2$, is an adjustment to the profit term, $2U'\Pi$, that appears in ϕ^1. If the dealer restricts himself to a very short horizon, transactions appear very profitable because he sees only the fee he collects immediately. As his horizon is extended, derived utility is adjusted downward to reflect the uncertainty of maintaining that profitability. The fourth term is a similar adjustment for the compound effect of Π and total portfolio return $r_W W$.

The last term is an adjustment for transactions uncertainty. Transactions uncertainty is not undesirable *per se*. It is undesirable only because it affects return uncertainty. Therefore σ_I^2 is part of this term. The expression, $\alpha/2$, is the expected arrival rate of purchases or sales at the dealer's a, b which maximizes expected profits (see footnote 7). When viewed in the context of a

[8] Since we are examining a short term interval, the change in dealer wealth is small so that he will not be drawn into the unreasonable range of the quadratic utility function.

[9] The procedure for deriving (23) is as follows. Take the derivative of (19),

$$J_{\tau\tau} = LJ_\tau + f_\tau^S + f_\tau^B,$$

where f^S and f^B are the second and third terms respectively of (19) and are functions of time From (20),

$$(J_{\tau\tau} | \tau = 0) = 2\phi^2, \qquad (J_\tau | \tau = 0) = \phi^1, \qquad \text{which is given by (22)}$$

Thus,

$$2\phi^2 = L\phi^1 + (f_\tau^S | \tau = 0) + (f_t^B | \tau = 0)$$

Taking the derivatives, f_τ^S and f_τ^B, is is laborious The steps are available by writing the authors.

[10] $L^2U = U''(rF + r_I I + r_Y Y)^2 + U'(r^2F + r_I^2 I + r_Y^2 Y) + U''(\sigma_I^2 r_I I^2 + \sigma_{IY} r_I IY + \sigma_{IY} r_Y IY + \sigma_Y^2 r_Y Y^2) + U''(r_I \sigma_I^2 I^2 + (r_I + r_Y)\sigma_{IY} IY + r_Y \sigma_Y^2 Y^2) + U''(\frac{1}{2}\sigma_I^4 I^2 + \sigma_{IY}^2 IY + \frac{1}{2}\sigma_Y^4 Y^2)$

small time interval, $\alpha/2$ is the probability of a purchase or sale (i.e., $(\alpha/2)\mathrm{d}t$ is the probability). Thus for a stock in which the probability of a transaction is large, the probability of incurring return risk (the value of which is $\sigma_I^2 Q^2$) is also large. Since $U'' < 0$, this term has a negative effect on J.

5. Dealer's pricing strategy

Having ϕ^0, ϕ^1 and ϕ^2, we have the solution for J as given by (20); and we can specify how the dealer's bid and ask prices depend on the return characteristics of the stock, the stochastic demand for dealer services and the dealer's elementary utility function. We describe the dealer's pricing strategy by specifying the dealer's spread, s, and a price adjustment variable, d. The spread is defined as $s=a+b$. (Since $p=1$, this is the proportional spread.) We show first that the spread is largely independent of the dealer's inventory position and depends on fundamental characteristics of the stock and the dealer. Second, the dealer encourages public customers to trade in the direction he desires by changing the adjustment variable which is defined as $d=a-b$, not by changing the spread. With the help of fig. 3 one can see that $d=0$ whenever the mid-point of the spread is at p, $d<0$ whenever the mid-point is less than p (and the dealer is relatively more anxious to sell than to buy), $d>0$ whenever the mid-point is greater than p (and the dealer is relatively more anxious to buy than to sell).[11] Given s and d, the selling fee and buying fee can always be derived according to

$$a=(s+d)/2, \qquad b=(s-d)/2.$$

6. The bid–ask spread

From (18a) and (18b) the dealer's spread is

$$s=\alpha/\beta+(J-SJ)/2SJ_F Q+(J-BJ)/2BJ_F Q. \tag{24}$$

The first term in (24) is the spread which maximizes expected revenues per share from sale and purchase transactions from the symmetric linear demand functions. The more inelastic the demand for dealer services the larger the spread.[12] The remaining terms are 'risk premiums' for a sale and purchase transaction, which assume optimal dealer behavior in the future. The sum of the risk premiums is required because the dealer sets the spread without knowing whether the next transaction will be a purchase or sale.

[11] Smidt (1979) has called the mid-point of the dealer's spread $(p+d)$, the market price, and he has called our p the equilibrium price. He views the dealer as adjusting his inventory by altering the difference between 'market price' and the 'equilibrium price', i.e., by altering our d.

[12] The papers of Garman (1976) and Amihud and Mendelson (1979) are concerned with this term or the relatea expression for profits.

Substitution of the solution for J into (24) yields the following optimal dealer spread:

$$s=\frac{\alpha}{\beta}-\tfrac{1}{2}Z(Q/W)\sigma_I^2\tau+\tfrac{1}{2}Z(Q/W)\sigma_I^2[(r_I-r+G_I)+Z(r_W+2\Pi/W)]\tau^2, \tag{25}$$

where $Z=U''W/U'$, the coefficient of relative risk aversion, and $G_I=r_I+\frac{1}{2}\sigma_I^2$, which is the instantaneous growth in the variance of I. The spread equation actually omits certain small second order terms. The full equation and the procedure for deriving it are in a mathematical appendix available from the authors.

We can now see what factors determine the spread and how the spread is changed as the horizon of the dealer is lengthened. When $\tau=0$, the dealer does not look beyond the current moment and is concerned only with the fee he can collect from a purchase or sale. Thus only the risk neutral, spread, α/β, already discussed above, is relevant.

At the time horizon increases, the second term of (25), which is the first-order risk adjustment, is relevant; and the spread is higher since all variables are positive. This term is one half the risk premium term derived in Stoll (1978a) where a one period framework was relied upon. The difference arises from the fact that we have maximized expected utility of a monopolistic dealer wheras in Stoll (1978a) demand conditions are not considered, and bid and ask prices are determined that maintain the dealer's original expected utility of terminal wealth. As was the case in Stoll (1978a), the first-order risk adjustment term does not depend on the dealer's inventory. Rather it depends on factors which are relatively stationary through time, i.e.:

(1) The dealer's attitude toward risk as reflected in the coefficient of relative risk aversion, Z.

(2) The relative value of the transaction for which the bid–ask quotation is made, Q. (Recall $p=1$.) Larger transactions require larger spreads.

(3) The risk of the stock as measured by the instantaneous variance, σ_I^2. Note that as far as the spread is concerned 'total risk' not just 'systematic risk' matters. With respect to base wealth, an efficient portfolio, the risk of inventory is measured by the covariance between the return on inventory and the return on base wealth, σ_{IY}. However that risk is offset by the expected excess return, r_I-r, and therefore does not appear in the final expression, (25). Because the dealer moves to a sub-optimal portfolio, the variance matters with respect to that fraction of his wealth, Q/W, which exceeds the optimal investment in the stock.

In fact, inventory position does not appear anywhere in (25) (except as it affects r_W). The reason is that the dealer adjusts the bid and ask prices equally whenever there is an inventory change. Thus when he acquires shares, he lowers his bid price (raises b) to discourage additional sales to him; and he lowers his ask price (lowers a) to encourage additional purchases from him. In other words he places himself on the margin where he is indifferent to a public sale or purchase because the higher compensation derived from a sale to him is equal in utility to the lower risk (*cum* smaller compensation) from a purchase from him.

Lengthening the dealer's time horizon makes the last term of (25) relevant and causes the spread to be still larger since the term is positive. For a given time horizon, one might expect the spread to be less under our multiperiod model than under the one period model of Stoll (1978a) since the dealer has an opportunity to adjust prices as his inventory changes. However, this is not the case. In a one-period model, one transaction occurs; and the dealer holds the risky position to the horizon. In our multiperiod model, additional transactions as well as dealer price adjustments may occur. The uncertainty of these additional transactions (which increases the uncertainty of the dealer's portfolio return and profits) more than offsets the risk reduction the dealer chooses to achieve by changing his prices. The single period approximation simply ignores these additional risks.

As in the case of the first-order risk adjustment, the second-order risk adjustment depends on the dealer's attitude toward risk, the transaction value, and the stock's return variance, which are multiplied by the bracketed term. The first term inside the brackets is a risk premium plus the growth in the variance of I. The second term is the coefficient of relative risk aversion multiplied by the return from two sources — the expected return on his total portfolio (r_W) and expected dealer profits expressed as a fraction of his wealth ($2\Pi/W$). Since the expected returns and growth in variance of I are likely to be small relative to the dealer's horizon, the principal effect comes from $2\Pi/W$: the larger the monopoly profits today the more the dealer stands to lose tomorrow, and therefore the larger the second-order risk adjustment.

Probably more surprising than the terms which enter (25) are the terms that do not enter. We had expected the transactions processes to play a more explicit role. They do not enter at all in the first-order risk term. They enter the second-order risk term through Π, but this is simply an adjustment for the expected growth in dealer wealth from his monopoly profits. There is no term for the variance of the transactions process. The absence of such a term in the spread equations is due to two factors. First, in this model transactions uncertainty is not undesirable *per se* because there is no direct adverse effect of a below average number of buyers or sellers. We have not assumed a minimum cost of doing business, for example. Transactions uncertainty is undesirable only because it increases the uncertainty of the return on the dealer's portfolio by making uncertain how large an unbalanced position must be held. The effect is seen in (23) in the term $(\alpha/2)\sigma_I^2 Q^2$. Second, the reason that this term does not appear in (25) is analogous to the reason that inventory does not appear. For any I, the dealer sets his bid and ask price so that he is indifferent on the margin as to whether the next transaction is a purchase or sale. This is true whatever is the probability of the next purchase or sale even if those probabilities were to differ because of asymmetric demand. As a result the term, $(\alpha/2)\sigma_I^2 Q^2$, does not enter the spread. Just as current inventory does not matter for the spread, the expected change in inventory does not affect the spread.[13]

7. Price adjustment

On the assumption of symmetric demand, the price adjustment variable derived from (18a) and (18b) is

$$d=\tfrac{1}{2}((J-SJ)/SJ_F Q)-\tfrac{1}{2}((J-BJ)/BJ_F Q). \tag{26}$$

The principal function of d is to promote adjustment of the dealer's inventory when it is temporarily thrown out of balance by the random arrival of transactions. Substituting the solution for J into (26) yields

$$d=-Z(I/W)\sigma_I^2\tau-Z(I/W)\sigma_I^2[r_I-r+G_I+Z(r_W+2\Pi/W)]\tau^2, \tag{27}$$

where certain small terms have again been neglected. The full equation is in a mathematical appendix available from the authors.

Eq. (27) is identical to (25) except for the fact that $2I$ replaces Q. Whereas the spread depends on the size of the transaction yet to come (Q), the price adjustment depends on the inventory which has been acquired (I). Eq. (27) is an inventory response equation that specifies the price adjustment variable be negative (positive) when inventory is positive (negative). When $d<0$, both the bid price and ask price are 'low' so as to encourage dealer sales and discourage dealer purchases, thereby reducing the dealer's inventory. When $I<0$, $d>0$ to encourage dealer purchases and discourage dealer sales. The degree of price response to an inventory change depends on the same factors determining the size of the spread — dealer's risk aversion, variance of the stock, and, in the second-order term, future increases in wealth which will be at risk.

[13] One might ask whether transactions uncertainty would be more critical if one introduced serial dependence in the process for purchases and sales. Postitive serial dependence in purchases and in sales would be like increasing the transaction size Q, and would therefore increase the spread Negative serial dependence would reduce the spread Whether any additional effects would enter is not at all obvious, and we have not worked out this case.

8. Asymmetric demand

Asymmetric demand arises if the dealer's perception of the true price, p, differs from the public's perception, p^*. For example if $p^* > p$, the expected number of dealer sales at p exceeds the expected number of dealer purchases at p as in fig. 1. Suppose the demand curves for dealer sales and purchases differ both in their slope and intercept,

$$\lambda_a = \alpha_a - \beta_a a, \tag{15c}$$

$$\lambda_b = \alpha_b - \beta_b b. \tag{15d}$$

Asymmetric demand alters the spread equation by changing the first term in (24) or (25) from α/β to $(\alpha_a/2\beta_a + \alpha_b/2\beta_b)$. However, if this term — the risk neutral spread — and expected dealer profits remain the same under symmetric as under asymmetric demand, there is no additional effect on spread through the risk premium terms in (25).[14] For example suppose that $\alpha_a = \alpha_b = 100$ initially, that α_a rises to 120 and α_b falls to 50, and that the appropriate changes occur in β_a and β_b to hold constant expected profits and the risk neutral spread. Then there is no additional effect through the risk terms despite the fact that the expected number of public sellers at p now exceeds the expected number of public buyers at p ($\alpha_a = 120$ vs $\alpha_b = 50$). This is because the dealer knows today that he can follow a price adjustment strategy in the future that makes him indifferent to this asymmetry.

What is crucially affected by asymmetric demand is the price adjustment variable, d. Just as the dealer adjusts prices if current inventory is non-zero so he adjusts prices for expected imbalances in demand which could drive future inventory to undesirable levels and thereby increase the dealer's risk to undesirable levels. Let d^1 denote the price adjustment under asymmetric demand. Following the procedure used in deriving (27), except that we now assume (15c) and (15d), gives

$$d^1 = [\alpha_a/2\beta_a - \alpha_b/2\beta_b] - \tfrac{1}{2}Z(Q/W)\sigma_I^2((\alpha_a - \alpha_b)/2)\tau^2 + d, \tag{28}$$

where d is given by (27).

The first term is simply the price adjustment necessary to maximize expected profits. This may require inventory accumulation because the stock is a good buy ($p > p^*$) or inventory reductions because the stock is a poor buy ($p < p^*$).[15]

[14] The term 2Π is now defined as $\alpha_a^2/4\beta_a + \alpha_b^2/4\beta_b$. See footnote 7. We consider only changes for which this term remains constant because we are not analyzing changes in the monopoly power of the dealer.

[15] This term may be zero even in the case of asymmetric demand if the demand curves are displaced so that profits continue to be maximized at the same prices. This is the case for a rotation that maintains the elasticity of demand; for example $\alpha_a = 2\alpha_b$ and $\beta_a = 2\beta_b$.

However, an expected imbalance in orders significantly raises the dealer's portfolio risk, and this requires a change in d by the amount of the second term of (28). The expected dollar inventory change is $((\alpha_a - \alpha_b)/2)Q$, and the relative inventory risk for which the dealer must adjust his price is therefore $((\alpha_a - \alpha_b)/2)(Q/W)\sigma_I^2$. If the probability of a sale exceeds that of a purchase ($\alpha_a > \alpha_b$), the dealer raises d^1 to discourage public sellers and encourage public buyers; and the converse is true if the probability of a purchase exceeds that of a sale. As in the case of the adjustment to current inventory, the magnitude of the change in d^1 depends on the instantaneous variance of the stock and the dealer's attitude toward risk.

These price adjustments do not eliminate inventory accumulation or reduction because the source of this inventory change is the favorable or unfavorable price of the stock in the opinion of the dealer. The price adjustment compensates the dealer for the loss of portfolio diversification and change in portfolio risk resulting from the additions to or reductions from inventory. If the dealer always sets $p = p^*$, where p^* is defined as the price at which the expected demand for dealer sales is equal to the expected demand for dealer purchases, $\alpha_a = \alpha_b$; and the case of asymmetric demand would not arise.[16]

9. Numerical illustration

Consider a dealer with a planning horizon of $\tau = 0.04$ of a year, approximately two weeks. For simplicity assume he faces symmetric demand. What will be his spread and price adjustment for the following assumed values of the exogenous variables?

$\alpha = 2000$ per year. Expected number of purchase transactions plus expected number of sale transactions at optimal dealer fee assuming symmetric demand.[17]

$\beta = 100{,}000$ per year. Determined to give risk neutral spread of $\alpha/\beta = 0.02$.

$Z = -U''W/U' = 2$. Coefficient of relative risk aversion, assumed constant.

[16] The problem then is how to determine expected demand One possibility is to base this expectation on recent transactions thereby causing $p = p^*$ to be based on the evolution of transactions.

[17] Under risk neutral pricing and symmetric demand the expected arrival rate of purchases or sales is $\alpha/2$. The number chosen is based on Smidt (1979) who found six of twelve NYSE stocks had fewer than 6000 transactions per year and on the assumption that the specialist participated as a dealer in one third of those transactions.

$\sigma_I^2 = 0.50$ per year.[18]

$r_I = 0.20$ per year, $r_Y = 0.22$ per year, $r = 0.10$ per year.

$\sigma_I^2 = 0.06$ per year, $\sigma_{IY} = 0.05$ per year.

We also assume the dealer begins anew from alternative initial positions on which he has not earned profits so that $F = -I$ and $W = Y$.

On the basis of these assumed values (25) and (27) now become

$$s^1 = \tfrac{1}{2}(Q/W)\tau + \tfrac{1}{2}(Q/W)[0.99 + 0.2I/W + 40Q/W]\tau^2, \qquad (25')$$

$$d = -(I/W)\tau - (I/W)[0.99 + 0.2I/W + 40Q/W]\tau^2, \qquad (27')$$

where $s^1 = s - \alpha/\beta$. The risk neutral spread, $\alpha/\beta = 0.02$ in this illustration, is invariant to the variables of interest to us, and we look only at the risk premium terms. It is obvious from the equation that, under the assumption of constant relative risk aversion, only the ratios, Q/W and I/W, are important. Values of s^1 and d for alternative values of these ratios are given in table 1. In the equation, the spread depends on I because the level of I affects portfolio return, r_W, in (25). However, as the table shows, this effect is less than one basis point and our earlier assertion that for practical purposes the spread is independent of inventory position is correct.

Table 1

Price adjustment (d) and risk premium component of spread (s^1), stated in percent, by transaction value as a fraction of wealth (Q/W) and inventory value as a fraction of wealth (I/W) Dealer's horizon is assumed to be $\tau = 0.04$ years

		I/W								
Q W		−0.40	−0 20	−0 10	−0 05	0.0	0 05	0 10	0 20	0 40
0 00	s^1	0.000	0.000	0 000	0 000	0.000	0.000	0.000	0.000	0.000
	d	1.668	0.830	0.415	0 208	0.000	−0 208	−0 416	−0 833	−1.668
0 05	s^1	0.112	0 112	0 112	0.112	0 112	0 112	0 112	0.112	0 112
	d	1.786	0 894	0 447	0.224	0.000	−0 224	−0 448	−0 897	−1 796
0.10	s^1	0 239	0.239	0.240	0.240	0 240	0.240	0 240	0 240	0 240
	d	1 914	0.958	0 379	0 240	0.000	−0.240	−0 480	−0.961	−1.924
0 20	s^1	0 542	0 543	0.543	0.543	0 544	0.544	0 544	0 544	0 545
	d	2 170	1.086	0.543	0 272	0 000	−0.272	−0 544	−1 089	−2 180
0.40	s^1	1 341	1.342	1.343	1 343	1.343	1.344	1 344	1.345	1 346
	d	2 682	1.342	0 671	0.336	0.000	−0 336	−0 672	−1 345	−2 692

[18] The annual return variance is on the high side in part to capture the special risk of the dealer from informational trading, which is not explicitly modeled here. By standing ready to trade at bid or ask price, the dealer incurs losses when investors have information Since the dealer's position tends to increase just before a price decline and decreases just before a price increase, his risk is greater than implied by the underlying variance of the stock's true return

It is evident from the table that for a wealthy individual the risks of making a market in a single stock can be small. For example for $Q/W=0.05$ the risk premium in the spread is only 0.122 percent ‘for values of I/W between -0.40 and $+0.40$) or \$6.10 on a \$5000 transaction. This implies a proportional buying or selling fee net of monopoly profits of only 0.061 percent or \$3.05. We expect that for certain classes of dealers, such as certain NYSE specialists, block positioning firms, or underwriters, the ratio, Q/W, may be considerably higher. And in such cases the risk premium term in the spread can increase significantly as the table shows. An obvious recommendation for improving the quality of securities markets is to increase the amount of wealth devoted to market making either by raising wealth of the dealer or by having more dealers. However one cannot compel an increase in a dealer's wealth, and institutional restrictions (as on the NYSE) as well as clerical and informational economics of scale may limit the number of dealers in a stock.

The values in table 1 depend on the time horizon of $\tau=0.04$. Doubling τ to 0.08 of a year would raise the spread (for $Q/W=0.05$) to 0.248 percent or \$12.40 on a \$5000 transaction. The spread more than doubles because the second-order term becomes relatively more important. Nevertheless its contribution to s^1 is only 0.048 percent, which implies that the degree of approximation achieved by eq. (10) is still adequate. While the variance of inventory return, $\sigma_I^2\tau$, increases as τ and the variance of inventory due to transactions, $(\lambda_a+\lambda_b)\tau$, increases as τ, the combined effect on portfolio risk, which appears in the second-order term increases faster than τ.

The table is helpful in understanding the behavior of bid and ask prices implied by our model. Let a prime on a variable indicate the risk premium component of the variable. Now consider a particular transaction size, say $Q/W=0.05$, where $s^1=0.00112$. At $I=0$, $d=0$, and

$$p_a^1=p+a^1=p+(s^1+d)/2=1.00056,$$

and

$$p_b^1=p-b^1=p-(s^1-d)/2=0.99944.$$

Suppose the dealer acquired \$5000 of inventory so that $I/W=0.05$. He would lower his prices by setting $d=-0.00224$ so that

$$p_a=p+(s^1+d)/2=0.99944,$$

and

$$p_b=p-(s^1-d)/2=0.99832.$$

The dealer has lowered bid and ask prices by the value of the spread, which, in the absence of monopoly profits, causes not only his bid price but also his ask price to fall below his opinion of the true price. Because the dealer has monopoly power it is unlikely that the ask price will in fact fall below p. Noting that $s=s^1+\alpha/\beta$ and $\alpha/\beta=0.02$, the actual ask and bid price corresponding to the previous example when $I/W=0.05$ are

$$p_a=p+(s+d)/2=1.00944,$$

and

$$p_b=p-(s-d)/2=0.98832.$$

If, by chance, no additional transaction occurs before the horizon date, the model assumes the dealer liquidates his inventory at p, thereby yielding an expected fee of b. If additional transactions occur that cause ending inventory to be zero, the dealer earns no risk premium because in effect he has borne no risk. Of course, prior to the horizon date, there is uncertainty about his ending inventory which increases the dealer's portfolio risk, and this causes him to demand a risk premium today.

10. Demand conditions and the decision to become a dealer

Our model is perhaps most applicable to the NYSE where each stock has only one dealer — the specialist. Consider the question of whether a newly listed stock would always find a specialist willing to make a market in the stock. That decision depends on the demand conditions and on the risk of being a dealer. The demand conditions in turn depend on the fundamental desires of investors to trade the stock and on the alternative sources of immediacy. There may be competing markets which reduce α and increase β to the specialist. Furthermore the demand for dealer services depends on the ease with which investors could 'make their own immediacy' by using limit orders and borrowing or lending. For example instead of selling immediately to the specialist at the specialist's p_b an investor could enter a higher limit price and borrow funds while he waits for his transaction to be executed. The cost to investors of these competing sources of immediacy influence the elasticity of demand for dealer services.

In terms of our model an individual will not become a dealer if the spread he must set to cover his risk is so large as to cut off the demand for his service. Under symmetric demand the spread required by the dealer is given by (25). Starting from a zero inventory position the spread at which the expected arrival of transactions is zero is $2\alpha/\beta$ (see footnote 7). Thus the condition for being a dealer is that the required spread be less than the spread at which trading is zero,

$$\begin{aligned}2\alpha/\beta > {} & \alpha/\beta+\tfrac{1}{2}Z(Q/W)\sigma_I^2\,\tau \\ & +\tfrac{1}{2}Z\,(Q/W)\sigma_I^2[(r_I-r+G_I)+Z\,(r_W+2\Pi/W)]\tau^2. \qquad (29)\end{aligned}$$

This condition can always be satisfied for transaction size, Q, small enough. But suppose there is some minimum Q the dealer must be ready to trade. Then the condition can be violated if α, the transactions arrival rate when the dealer's fee is zero, is sufficiently small. For example when $\alpha=200$, $Q/Y=0.1$, $z=2$, $\sigma_I^2=0.5$ per year, $r=0.10$ per year, $r_Y=0.20$ per year, $\beta=100{,}000$, and $\tau=0.04$, the dealer would not make a market, for in that case expected profits would not offset the risk. It is also clear that condition (29) is less likely to be satisfied the larger an individual's Z or the smaller his wealth. Regulations to increase the transactions size for which a bid–ask quotation must be honored would reduce the willingness of individuals to become a dealer. Since the NYSE specialist also earns brokerage on limit orders he executes, one may see an individual declare his willingness to be a specialist without at the same time being willing to act as a dealer. In such circumstances the provision of immediacy by the specialist is likely to be particularly poor.

11. Summary

We have considered the optimal behavior of a single dealer in a single stock who is faced with a stochastic demand for his services (modeled by a continuous time Poisson jump process) and faces return risk on his stock and on the rest of his portfolio (modeled by diffusion processes). As time unfolds and transactions occur, the dealer is able to set his bid price and ask price relative to his opinion of the 'true' price of the stock so as to maximize the expected utility of terminal wealth. This stochastic dynamic programming problem is solved for the dealer's derived utility function; and using this solution, we are able to specify how the dealer sets the bid and ask price at any moment of time as a function of the state in which he finds himself. Indeed the solution could be programmed on the computer.

We show that the spread, the difference between the ask price and bid price, is given by a risk neutral spread that maximizes expected profits for the given stochastic demand functions plus a risk premium that depends on transaction size, the return variance of the stock and the dealer's attitude toward risk. Consistent with the one period model of Stoll (1978a), the spread does not depend on the dealer's inventory position. However dealer price adjustment does depend on inventory. When inventory increases both bid price and ask price decline, and the converse is true when inventory decreases.

Despite the dealer's ability to adjust prices in response to inventory changes as time unfolds, the dealer's risk is greater than implied by the one period model of Stoll (1978a). This results from the fact that the uncertainty of the demand to trade with the dealer is not totally eliminated or offset by the dealer's pricing strategy.

Other issues investigated are the case of asymmetric demand for trading with the dealer and the conditions under which an individual chooses or refuses to make a market in a stock. In an inactive stock, when the dealer is required to trade a minimum amount, the expected profit from trading may not be enough to offset the risk.

Although couched in terms of a dealer, our model can apply to any individual who is determining his reservation price for a stock as a function of his opinion of the stock's value and the degree to which the stock will cause his portfolio to be unbalanced. Thus the placement of limit orders by individuals could, for example, be modeled by our approach.

References

Amihud, Y. and H. Mendelson, 1980, Dealership market: Market making with inventory, Journal of Financial Economics 8, no. 1, 31–54.

Benston, G. and R. Hagerman, 1974, Determinants of bid–asked spread in the over-the-counter market, Journal of Fmancial Economics 1, no. 4, 353–364.

Blattberg, R. and N. Gonedes, 1974, A comparison of the stable and student distribution as statistical models for stock prices, Journal of Business 47, April.

Clark, Peter, 1973, A subordinated stochastic process model with finite variance for speculative prices, Econometrica 41, Jan.

Cohen, K.J., S.F. Maier, R.A. Schwartz and D.K. Whitcomb, 1978a, The returns generation process returns variance and the effect of thinness in securities markets, Journal of Finance 33, March.

Cohen, K.J., S.F. Maier, R.A. Schwartz and D.K. Whitcomb, 1978b, Limit orders, market structure, and the returns generation process, Journal of Finance 33, June.

Cohen, K.J., S.F. Maier, R.A. Schwartz and D.K. Whitcomb, 1979, Market makers and the market spread: A review of recent lıterature, Working paper, Aug.

Copeland, T., 1976, A model of asset trading under the assumption of sequential ınformation arrival, Journal of Finance 31, Sept.

Davis, M.A.A., 1977, Linear estimation and stochastic control (Chapman and Hall, London, and Wiley, New York).

Epps, T., 1976, The demand for broker's services: The relation between security trading volume and transaction cost, Bell Journal of Economics 7, Spring.

Garbade, K. and W. Silber, 1979, Structural organization of secondary markets: Clearing frequency, dealer activity and liquidity risk, Journal of Finance 34, June.

Garman, Mark, 1976, Market microstructure, Journal of Financial Economics 3, no. 3, 257–275.

Goldman, M.B. and A. Beja, 1979, Market prices vs equilıbrıum prices: Returns variance, serial correlation and the role of the specialist, Journal of Finance 34, June.

Goldman, M.B. and H.B. Sosin, 1979, Informatıon disseminatıon, market effıcıency and the frequency of transactions, Journal of Financial Economics 7, no. 1, 29–61.

Ho, T. and H. Stoll, 1980, On dealer markets under competition, Journal of Finance 35, May.

Mildenstein, E. and H. Schleef, 1980, The optimal pricing policy of a monopolistic marketmaker in the equity market, Working paper (University of Oregon, Eugene, OR).

Oldfield, G., R. Rogalski and R. Jarrow, 1977, An autoregressive jump process for common stock returns, Journal of Financial Economics 5, no. 3, 389–418.

Schwartz, R. and D. Whitcomb, 1977, The time–variance relationship: Evidence on autocorrelation in common stock returns, Journal of Finance 32, March.

Smidt, Seymour, 1979, Continuous vs. intermittent trading on auction markets, Paper presented at the Westerm Finance Association Meetings, June 22, 1979.

Stoll, Hans R., 1978a, The supply of dealer services in securities markets, Journal of Finance 33, Sept.

Stoll, Hans R., 1978b, The pricing of security dealer services: An empirical study of NASDAQ stocks, Journal of Finance 33, Sept.

Tinic, Seha, 1972, The economics of liquidity services, Quarterly Journal of Economics 86, Feb.

Tinic, Seha and R. West, 1972, Competition and the pricing of dealer services in the over-the-counter market, Journal of Financial and Quantitative Analysis 7, June.

Westerfield, R., 1977, The distribution of common stock price changes. An application of transactions time and subordinated stochastic models, Journal of Financial and Quantitative Analysis, Dec.