16
The Determinants of the Bid–Ask Spread

Another issue brought to the fore by the crisis is the need to better understand the determinants of liquidity in financial markets. The notion that financial assets can always be sold at prices close to their fundamental values is built into most economic analysis…

(Ben Bernanke)

As we discussed in Chapter 1, organising a market to ensure fair and orderly trading is by no means a trivial task. As we also discussed in Chapter 1, transactions can only take place if some market participants post binding quotes to the rest of the market, in the sense that they specify prices at which they agree to buy or sell a specified quantity of an asset. By posting these quotes, liquidity providers put themselves at risk, because liquidity takers can decide whether or not they want to accept these offers to trade – and will only do so if they believe that the price is favourable. Liquidity providers are therefore exposed to a systematic, adverse bias: while some trades are uninformed and innocuous, other trades may be informed and be followed by large price moves in the direction of the trade, to the detriment of the liquidity provider.[1]

Given this seemingly unfavourable position, why do any market participants provide liquidity at all? The answer is that many liquidity-provision strategies, including the popular strategy of market-making, can be profitable in the long run because a large fraction of trades are in fact non-informed (or very weakly informed). As we noted in Section 1.3.2, the fundamental consideration for implementing these strategies in the long run is the balance between the mean size of the bid–ask spread and the mean strength of adverse impact.

[1] As noted by Perold, A. F. (1988). The implementation shortfall. Paper versus reality. *The Journal of Portfolio Management*, 14(3), 4–9: "*You do not know whether having your limit order filled is a blessing or a curse – a blessing if you have just extracted a premium for supplying liquidity; a curse if you have just been bagged by someone who knows more than you do.*"

In older financial markets, only a select few market participants were able to act as market-makers. Due to the large spreads that were typical of these markets, market-making was a highly profitable business for these individuals. In modern electronic markets, by contrast, any market participant can act as a market-maker. As we discuss in the chapter, this important change has made market-making a highly competitive business that is typically only marginally profitable.

In this chapter, we introduce and study some simple models that help to make our previous discussions of market-making and the bid–ask spread more precise. As we will see, modern trade-and-quote data confirms the existence of a close correspondence between impact, volatility and the bid–ask spread, enforced by competition between liquidity providers.

16.1 The Market-Maker's Problem

Market-makers attempt to earn profit from exploiting the difference between the bid- and the ask-price using (primarily) limit orders. Understanding how to choose the value of the bid–ask spread is a question of paramount importance for a market-maker: choosing a value that is too small will leave a market-maker under-compensated for the **adverse selection** risk inherent in implementing the strategy; choosing a value that is too large will prevent the market-maker from conducting any trades, as other liquidity providers will offer trades at better prices.

We begin by examining one of the first models to formalise the issue of adverse selection in an LOB framework. This model was originally due to a paper from Glosten and Milgrom in 1985, and is close in spirit to the Kyle model that we discussed in Chapter 15. As in the Kyle model, one assumes that some market participants are informed, in the sense that they have access to private information about the price of the asset at some future **terminal time**. Other market participants are either uninformed (even if they may believe otherwise!) or trade for other reasons, without any view on the future price.

16.1.1 Break-Even Quotes

In the model, the informed and uninformed agents trade with a single market-maker, who we again call Bob. Bob chooses a bid-price b, at which he places a unit volume for purchase, and an ask-price a, at which he places a unit volume for sale. We assume that Bob chooses the values of b and a so as to ensure he has **no ex-post regrets**, in the sense that the true price of the asset (revealed after the trade) is on average equal to b if an agent sold (and Bob bought) and is on average equal to a if an agent bought (and Bob sold). This quote-setting rule allows Bob to break even on average. In other words, we assume that Bob is risk-neutral and that he does not increase the bid–ask spread to compensate for the variance and skewness of his payoff (see Section 1.3.2). This is justified in a competitive situation, which pushes any risk-compensating premium to small values.

In addition to the market-maker, the **Glosten–Milgrom model** assumes that the market is populated by a set of liquidity takers. Each liquidity taker i maintains a private (idiosyncratic) valuation $\widehat{p_i}$, which is a random variable. If i is an informed trader, then $\widehat{p_i}$ is positively correlated with the terminal value p_{F}; if i is an uninformed trader, then $\widehat{p_i}$ is independent of p_{F}. Each liquidity taker i buys or sells the asset depending on $\widehat{p_i}$ and on Bob's quotes. Specifically, if $\widehat{p_i} > a$, then agent i buys the asset from Bob at price a; if $\widehat{p_i} < b$, then agent i sells the asset to Bob at price b. If $b < \widehat{p_i} < a$, the liquidity taker does nothing.

Within this framework, the conditions for Bob to have no ex-post regrets are given by

$$\begin{aligned} a &= \mathbb{E}_{\mathrm{Bob}}[p_{\mathrm{F}}|\widehat{p} > a], \\ b &= \mathbb{E}_{\mathrm{Bob}}[p_{\mathrm{F}}|\widehat{p} < b]. \end{aligned} \tag{16.1}$$

Equation (16.1) says that conditional on a buy trade occurring at the ask, the expected future ask-price (from Bob's standpoint) is equal to the price paid by the buyer. Symmetrically, conditional on a sell trade occurring at the bid, the expected future bid-price (from Bob's standpoint) is equal to the price paid by the seller. In both cases, the expectation is taken with respect to the information available to Bob. As in the Kyle model (see Chapter 15), this is not the full set of information available in the market. If Bob knew the terminal value p_{F}, then his break-even condition would simply be $a = b = p_{\mathrm{F}}$, since he would not be subject to any adverse selection, and all trades would be noise. In general, however, $b < a$ since a sell order brings a negative piece of information while a buy order brings a positive piece of information.

16.1.2 A Model with Well-Informed Traders

The simplest model for Bob's view of the world is that a fraction of agents are perfectly informed and know the value of p_{F}, while others are uninformed and their expectation of the future price is symmetrically distributed around the current price p_0:

$$\mathbb{P}(\widehat{p}|p_{\mathrm{F}}) = \underbrace{(1-\phi)f(\widehat{p}-p_0)}_{\text{uninformed}} + \underbrace{\phi\delta(\widehat{p}-p_{\mathrm{F}})}_{\text{informed}}, \tag{16.2}$$

where $f(\cdot)$ is a certain symmetric distribution function, ϕ is the fraction of informed traders, and the δ-function reflects that informed traders perfectly forecast the terminal price.[2]

[2] In fact, adding some uncertainty around p_{F} (i.e. fattening the δ-function) does not change the qualitative conclusions of the model.

In order to set his quotes using Equation (16.1), Bob needs to estimate the distribution of p_F, given that someone trades. Bayes' rule allows him to compute

$$\mathbb{P}(p_\mathrm{F}|\widehat{p}) = \frac{\mathbb{P}(\widehat{p}|p_\mathrm{F})\mathbb{Q}_0(p_\mathrm{F})}{\int_{-\infty}^{+\infty} \mathrm{d}p_\mathrm{F}\, \mathbb{P}(\widehat{p}|p_\mathrm{F})\mathbb{Q}_0(p_\mathrm{F})}, \qquad (16.3)$$

where $\mathbb{Q}_0(p_\mathrm{F})$ denotes the prior distribution of p_F at time $t = 0$. As within the Kyle framework, we assume that Bob knows $\mathbb{Q}_0(p_\mathrm{F})$, but of course not the value of p_F itself. Bob can now use Equation (16.3) to compute his required conditional expectations (see Equations (16.1)):

$$a = \mathbb{E}_\mathrm{Bob}[p_\mathrm{F}|\widehat{p} > a] = \frac{\int_{-\infty}^{+\infty} \mathrm{d}p_\mathrm{F}\, p_\mathrm{F}\, \mathbb{P}(\widehat{p} > a|p_\mathrm{F})\mathbb{Q}_0(p_\mathrm{F})}{\int_{-\infty}^{+\infty} \mathrm{d}p_\mathrm{F}\, \mathbb{P}(\widehat{p} > a|p_\mathrm{F})\mathbb{Q}_0(p_\mathrm{F})},$$

$$b = \mathbb{E}_\mathrm{Bob}[p_\mathrm{F}|\widehat{p} < b] = \frac{\int_{-\infty}^{+\infty} \mathrm{d}p_\mathrm{F}\, p_\mathrm{F}\, \mathbb{P}(\widehat{p} < b|p_\mathrm{F})\mathbb{Q}_0(p_\mathrm{F})}{\int_{-\infty}^{+\infty} \mathrm{d}p_\mathrm{F}\, \mathbb{P}(\widehat{p} < b|p_\mathrm{F})\mathbb{Q}_0(p_\mathrm{F})}.$$

To give some flesh to these equations, we choose some specific form for the uninformed distribution f (in Equation (16.2)) and for $\mathbb{Q}_0$. Our choice, motivated by the simplicity of the resulting calculations, reads:

$$f(u) = \frac{e^{-|u|/w}}{2w}; \qquad \mathbb{Q}_0(p_\mathrm{F}) = \frac{e^{-|p_f - p_0|/\sigma}}{2\sigma},$$

where w is a parameter that controls the dispersion of the uninformed price expectations and σ is proportional to the volatility of the fundamental price. This choice of distribution allows us to obtain explicit results, but the resulting conclusions still hold qualitatively for a wide class of distributions.

By performing elementary integrals of exponential functions in the expressions for a and b, the final result reads:

$$a = p_0 + s/2, \\ b = p_0 - s/2, \qquad (16.4)$$

with

$$s = \frac{\phi(2\sigma + s)e^{s/w}}{\phi e^{s/w} + (1-\phi)e^{s/\sigma}}. \qquad (16.5)$$

Equation (16.5) is quite interesting and exhibits different regimes depending on the ratio w/σ, as we now detail.

16.1.3 An Orderly Market

In the case where $w \geq \sigma$, Equation (16.5) has a single solution, which, in the limit of a small fraction of informed traders (i.e. $\phi \to 0$) reads:[3]

$$s \approx 2\phi\sigma + O(\phi^2), \qquad (\phi \ll 1). \tag{16.6}$$

In other words, when uninformed traders have sufficiently dispersed signals with an amplitude w that is greater than the width σ of the informed signal, Bob should set s to be proportional to the product of the price volatility σ and the fraction of informed trades ϕ. Note that in the limit $\phi \ll 1$ one has $w \gg s$: uninformed agents trade (almost) unconditionally.

Since in reality the ratio of the bid–ask spread s to the daily volatility σ is very small (say, 0.01 to 0.1), Equation (16.6) suggests that the fraction of trades that are informed about daily price moves must indeed be small. In fact, the value of σ for informed traders increases with prediction horizon T as the price volatility itself (i.e. as $\sqrt{T}$). The corresponding fraction of informed traders (or the quality of their information; see Equation (16.9)) must therefore be bounded from above by $1/\sqrt{T}$ for the bid–ask spread to remain bounded, independent of the horizon of informed traders. In a nutshell, the fraction of the volatility captured by informed traders must decrease with the time horizon.

It is also interesting to consider the **skewness** ς of the distribution of future price changes. Conditional on a buy trade occurring, the updated distribution of price changes is $\mathbb{P}(p_{\rm F}|\widehat{p} > a)$. In this case, what is the third moment ς of the distribution of $p_{\rm F} - a$? To first-order in ϕ (assumed to be small):

$$\varsigma \approx \varsigma_0 + \phi \frac{\int_0^\infty \mathrm{d}p_{\rm F}\,(p_{\rm F} - p_0)^3 \mathbb{Q}_0(p_{\rm F})}{\left[\int_{-\infty}^\infty \mathrm{d}p_{\rm F}\,(p_{\rm F} - p_0)^2 \mathbb{Q}_0(p_{\rm F})\right]^{3/2}},$$

where ς_0 is the unconditional skewness of price changes, and the second term is strictly positive. Hence, even for a symmetric unconditional distribution of price changes $\mathbb{Q}_0(p_{\rm F})$, observing a buy (respectively, sell) transaction should lead to a positively (respectively, negatively) skewed distribution of realised price changes $p_{\rm F} - a$.

One can in fact easily remove the distracting contribution of ς_0 by studying the skewness of the random variable

$$U := \mathbb{I}_{\widehat{p}>a}(p_{\rm F} - a) + \mathbb{I}_{\widehat{p}<b}(b - p_{\rm F}),$$

which simply represents the P&L of market orders. Again, to first-order in ϕ, it follows that $\varsigma(U) \propto \phi$. Figure 16.1 shows the (low-moment) skewness of U, estimated as the difference between the median and the mean (scaled by the r.m.s.), as a function of the time horizon T, and measured in market order event-time. We estimate p_F as the mid-price T market orders later,

$$\widetilde{U}(T) := \mathbb{I}_{\widehat{p}_t>a_t}(m_{t+T} - a_t) + \mathbb{I}_{\widehat{p}_t<b_t}(b_t - m_{t+T}). \tag{16.7}$$

For the small-tick stock (TSLA), the skewness is positive and decays roughly as $1/\sqrt{T}$, as predicted by the Glosten–Milgrom model. The skewness is however very

[3] This can easily be seen by assuming that s can be expanded as $c\phi + c'\phi^2 + o(\phi^2)$, then plugging this expression into Equation (16.5). Identifying the two sides of the equation order by order in ϕ gives Equation (16.6).

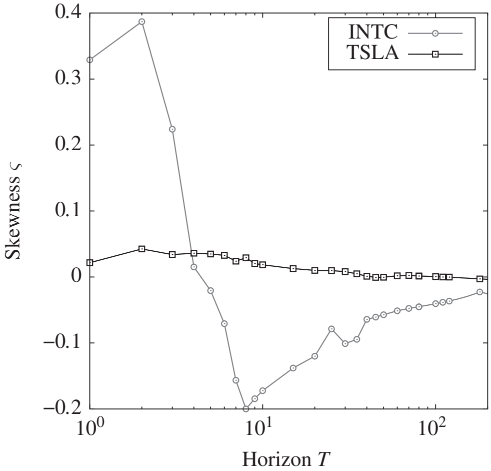

Figure 16.1. Skewness of $\widetilde{U}(T)$ (defined in Equation (16.7)), estimated with the low-moment difference between its median and its mean, as a function of the time horizon T, measured in market order event-time for INTC and TSLA.

small, even on short time scales, suggesting again that the fraction of informed trades is itself very small. For the large-tick stock (INTC), the skewness quickly becomes negative, in contradiction with the Glosten–Milgrom framework. In the case of large-ticks, however, we know that the bid–ask spread is bounded from below and the Glosten–Milgrom argument must be amended.

16.1.4 Market Breakdown

We now consider the more interesting case of Equation (16.5), where $w < \sigma$. In this case, uninformed agents do not predict very large price moves, so their signal rarely exceeds the bid–ask spread and they do not trade much when the spread is large. Figure 16.2 shows that a second, large-s solution appears in the interval $w^* < w < \sigma$, where w^* is a threshold that does not have a simple analytical form. Competition between market-makers enforces the solution with smallest s.[4] But when $w \downarrow w^*$, the large- and small-s solutions converge before they both disappear for smaller values of w.

The case with no solutions is particularly interesting: it suggests that the market breaks down because there is no longer any way for Bob to fix a bid–ask spread to break even. Put simply, the uninformed traders do not provide enough potential gains to compensate Bob for the adverse selection that he experiences from the informed traders, whatever the value of the spread he chooses. In the Kyle model, where traders send orders *before* knowing the transaction price, reducing the number of noise traders results in an increase of the market-impact parameter Λ,

[4] Still, one may expect in that case that the spread "hesitates" between the two solutions, leading to interesting regime shifts.

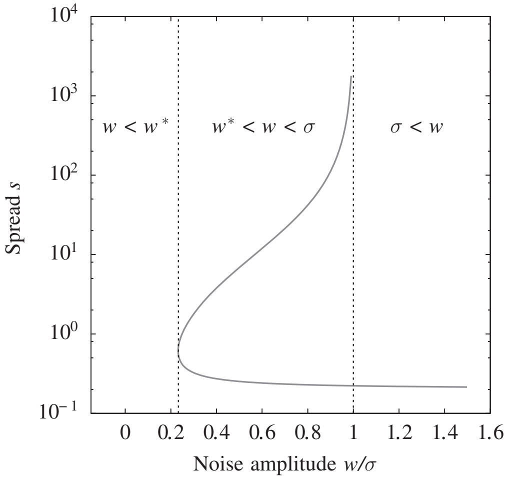

Figure 16.2. Numerical solution of Equation (16.5) for $\phi = 0.1$ and $\sigma = 1$ and different values of w. Note that there is a unique solution for $w > \sigma$, two solutions for $w^* < w < \sigma$ and no solutions for $w < w^* \approx 0.22\sigma$.

but without the market breaking down. In both cases, the presence of noise traders is crucial to ensure orderly trading.

This emergence of **market breakdown** is one of the most interesting features of the Glosten–Milgrom model. The effect would still exist even if informed traders had only partial information about the future price (i.e. when the δ-function in Equation (16.2) is replaced by a wider distribution). In fact, one can consider the case where all traders are partially informed, with some uncertainty about the true future price. For example, we might assume that Equation (16.2) is replaced by a logistic distribution

$$\mathbb{P}(\widehat{p}|p_{\mathrm{F}}) = \frac{\epsilon e^{-\epsilon(\widehat{p}-p_{\mathrm{F}})}}{\left(1+e^{-\epsilon(\widehat{p}-p_{\mathrm{F}})}\right)^2}, \tag{16.8}$$

where ϵ describes the amount of information on the future price that all traders (except the market-maker) have at their disposal. In the limit of weak information (i.e. $\epsilon \to 0$), solving Equation (16.8) leads to

$$s \approx \epsilon\sigma. \tag{16.9}$$

This solution is very similar to the one in Equation (16.6). Again, the solution disappears when ϵ exceeds a critical value ϵ_c, which corresponds to the case when market-making becomes impossible because the information available to traders is too precise.

In summary, markets can only operate smoothly if the number of uninformed traders is sufficiently high. When market-makers fear that other market participants have too much information, they are unable to fix any spread that enables them to break even, which can lead to a **liquidity crisis**.

16.1.5 A Dynamical Version of the Model: Quote Updates

The setting of the above discussion was entirely static. What happens after the first trade has occurred? Throughout this section, we assume that a solution for s always exists (i.e. $w > \sigma$) and that the fundamental price (which Bob attempts to guess from the order flow) is fixed in time.

Assume that a trade has taken place at the ask a_t at time t. Bob should now use the information that the trade has occurred to update his unconditional distribution of the future price p_F, so that at time $t+1$:

$$\mathbb{Q}_{t+1}(p_\text{F}) = \mathbb{P}_t(p_\text{F}|\widehat{p_t} > a_t).$$

The right-hand side of this expression is simply given by Equation (16.3), with $\mathbb{Q}_0$ replaced by $\mathbb{Q}_t$ to reflect that we now consider a dynamical model. This, in turn, allows Bob to fix the next bid- and ask-prices, again imposing a break-even constraint. The detailed analysis of the induced price dynamics is beyond the scope of this book, but one can show that the resulting sequence of trade prices is a martingale that converges to p_F at the terminal time when the number of steps goes to infinity, while the bid–ask spread exponentially decreases in time, because transactions gradually reveal the signal of the informed agents.

The **martingale** property (using public information available to the market-maker) is in fact quite simple to prove. Observe that

$$p_t = \begin{cases} a_t, & \text{if the trade at time } t \text{ is a buy,} \\ b_t, & \text{if the trade at time } t \text{ is a sell.} \end{cases}$$

From Equation (16.1), one has that immediately after each transaction t, it must hold that

$$\mathbb{E}_t[p_\text{F}] = p_t.$$

The expected transaction price at the next trade can therefore be written as:

$$\begin{aligned} \mathbb{E}_t[p_{t+1}] &= \mathbb{E}_t\Big[\mathbb{E}_{t+1}[p_\text{F}]\Big], \\ &= \mathbb{E}_t[p_\text{F}], \\ &= p_t, \end{aligned}$$

which is precisely the martingale property.

Is the **exponential spread decay** (in time) predicted by Glosten–Milgrom also observable in empirical data? Figure 4.2 indeed shows that the spread tends to decay as the day proceeds. In the Glosten–Milgrom framework, this means that overnight shocks introduce some uncertainty on the fundamental price, and that this uncertainty gets progressively resolved throughout the day as smart traders work out the consequence of these shocks. However, the plot suggests that the precise dynamics in real markets is more complex than the prediction of the model. In particular, an exponential fit of the decay is not adequate. Clearly, the idea of a fixed fundamental price p_F that the market strives to discover throughout the day is grossly over-simplified. In fact, the target price itself changes throughout the day, making the dynamics somewhat more intricate.

We conclude this section with a remark that is similar to the one that we made in the context of the Kyle model in Section 15.3. The dynamical update process that we have considered in this section requires Bob to have an unbiased estimate of the quantity of information available to informed traders at each instant of time. If Bob tends to overestimate the information content of trades (perhaps in an attempt to protect himself against selection bias), then his quote updates are not justified by the fundamental valuation of the asset. However, if other traders adjust to these new quotes and change their views on the price, self-fulfilling prophecies and excess volatility may emerge.[5]

16.1.6 Metaorders in the Glosten–Milgrom Framework

In the Glosten–Milgrom framework, the permanent impact of a trade is the expected value of the price, given that the specified trade has taken place. As we discussed in Section 10.5, however, many traders in real markets do not simply submit individual market orders, but instead seek to execute large metaorders. To minimise the impact of their actions, traders typically submit their child orders for a single metaorder over a period of several days. In the context of the Glosten–Milgrom model, it is interesting to measure the permanent impact of not just single market orders, but also of metaorders.

To address this case, we follow Farmer et al.[6] and make the additional assumptions that metaorder executions occur one-after-the-other (and therefore do not overlap), and that Bob is notified (or is able to infer) every time that a new metaorder execution begins.

Recall from Section 12.2 that a metaorder has a direction ε and total volume Q, and is split into child orders. We now assume that each child order has a volume equal to the lot size υ_0 (see Section 3.1.5). Let q denote the volume of the metaorder executed so far, and let $p(q)$ denote the price set by Bob at that point. Let $p_\infty(Q)$ denote the expected price after a metaorder of total volume Q has terminated and impact has relaxed.

As in the previous section, we assume that Bob sets his price such that each time he receives a market order, he has no ex-post regrets and breaks even on average. As we discussed in Section 16.1.5, this can be seen as a martingale condition for the price. In the present context, these two conditions read:

(i) **Martingale condition**: Bob's action makes the price process p a martingale. Therefore, when a volume q has been executed, the price is such that

$$p(q) = \mathbb{E}[p_\infty(Q) \mid Q \geq q]. \tag{16.10}$$

[5] For more on this scenario, see also Section 20.3.

[6] Farmer, J. D., Gerig, A., Lillo, F., & Waelbroeck, H. (2013). How efficiency shapes market impact. *Quantitative Finance*, 13, 1743–1758.

(ii) **Break-even condition**: Irrespective of the metaorder size Q, the price must be such that

$$\frac{1}{Q}\int_0^Q p(q)\mathrm{d}q = \mathbb{E}\left[p_\infty(Q)\right], \tag{16.11}$$

(where for simplicity we consider volumes and prices to be continuous, not discrete).

In words, the second condition simply stipulates that

$$\text{impact cost} = \mathbb{E}[\text{permanent impact}].$$

This is a metaorder-equivalent condition to the Glosten–Milgrom condition for single market orders, which reads

$$\text{half spread} = \mathbb{E}[\text{adverse selection}].$$

Together with the distribution of metaorder sizes $F(Q) := \mathbb{P}[q > Q]$, these two conditions allow one to determine the temporary impact $\Im(q) = p(q) - p_0$ and the permanent impact $\Im_\infty(Q) = p_\infty(Q) - p_0$ of a metaorder. Noting that the market-maker receives a volume flow that stops only when the metaorder is completed, one finds

$$\begin{aligned} p(q) &= \frac{1}{F(q)}\int_q^\infty \mathrm{d}Q f(Q) p_\infty(Q), \\ p_\infty(Q) &= \frac{1}{Q}\int_0^Q \mathrm{d}q p(q), \end{aligned} \tag{16.12}$$

where $f(q) := -F'(q)$. By multiplying each of these equations by the denominator term on the right-hand side, then taking derivatives (using the product rule), one finds a pair of coupled differential equations:

$$F(q)p'(q) - f(q)p(q) = -f(q)p_\infty(q), \tag{16.13}$$

$$p_\infty(q) + qp'_\infty(q) = p(q). \tag{16.14}$$

By substituting the expression for p_∞ obtained from Equation (16.13) into Equation (16.14), we arrive at the following first-order differential equation for $p(q)$:

$$q\left(\frac{F}{f}p'\right)' + \left(\frac{F}{f} - q\right)p' = 0.$$

This equation can be solved by introducing an auxiliary function

$$g(q) := \frac{F}{f}p'(q),$$

leading to

$$g(q) = \frac{A}{qF(q)},$$

where A is a constant.

Finally, using the fact that $\Im(0) = 0$, it follows that

$$\Im(q) = A\int_0^q \mathrm{d}q' \, \frac{f(q')}{q'F(q')^2}. \tag{16.15}$$

Hence, if metaorder volumes are distributed according to a power-law $F(q) \sim q^{-1-\gamma}$, where $0 < \gamma < 1$, then for large q, mean impact is asymptotically concave:

$$\Im(q) \propto q^\delta, \qquad \delta = \gamma. \tag{16.16}$$

The mean permanent impact then follows from the fair-pricing condition:

$$\Im_\infty(Q) \underset{Q\to\infty}{\approx} \frac{1}{1+\delta}\Im^{\text{peak}}(Q). \tag{16.17}$$

In the special case where $\gamma = 1/2$ (see Chapter 10), we recover an asymptotically square-root impact corresponding to $\delta = 1/2$. This is essentially the argument of Farmer et al. for a square-root impact law. In this case, the permanent impact is such that

$$\mathfrak{I}_\infty(Q) \approx \frac{2}{3}\mathfrak{I}^{\text{peak}}(Q).$$

According to the fair-pricing condition, if impact is square-root, then two-thirds of the peak impact should remain permanent.

The argument that we have presented in this section appears to be quite general, and is in fact very similar to the one in Section 13.4.5, where we studied impact in the context of the propagator model when trade signs arrive in sequences whose lengths are distributed according to a power-law. However, this theory suffers from several drawbacks. First, the relation between the impact exponent δ and the power-law exponent $1+\gamma$ for the distribution of metaorder sizes does not hold universally.[7]

Second, it makes the highly unrealistic assumptions that metaorders appear sequentially, in isolation, and that the market-maker can detect their start and end. Third, the model suggests that impact decays instantaneously to its asymptotic value after the metaorder terminates (see Section 13.4.5), whereas empirical data reveals that this behaviour is not routinely observed in real markets. Instead, impact undergoes an initially steep decay, then relaxes very slowly over a period that can span several days. Finally, it only holds in a large-Q regime, where the volume distribution does follow a power-law. For small volumes, its prediction deviates from the square-root law.

However, a strong point for the model is that empirical data suggests real market dynamics to be compatible with the prediction of a permanent impact equal to about two-thirds of peak impact. This two-thirds ratio appears to hold for metaorders that originate from some sort of information, rather than just for liquidity purposes (see Section 12.7 for recent papers on this point). Like the Glosten–Milgrom argument, according to which the bid–ask spread compensates for adverse selection, the fair-pricing argument is more general than the detailed set-up of the above model.

16.2 The MRR Model

Although inspiring, the Glosten–Milgrom model suffers from a problem: central to its formulation is the idea of a terminal time where the true price p_{F} is revealed, and at which people can transact as much as they want. The implicit idea (that also underlies Kyle's model; see Chapter 15) is that the market-maker's inventory can be fully transacted at the end of the day (say). The common lore is indeed to assume that one can trade large quantities at the market close without impacting the price, but this is totally unwarranted. Liquidating the market-maker's inventory incurs a potentially large impact cost and makes the Glosten–Milgrom break-even argument shaky.

Therefore, although the Glosten–Milgrom model is a useful starting point for thinking about how market-makers might choose to set the quotes that they

[7] See, e.g., Mastromatteo, I., Tóth, B., & Bouchaud, J.-P. (2014). Agent-based models for latent liquidity and concave price impact. *Physical Review E*, 89, 042805. In this paper, the LMF relation between the exponent of the order size distribution and the sign autocorrelation exponent γ was assumed. But in Donier, J., & Bonart, J. (2015). A million metaorder analysis of market impact on the Bitcoin. *Market Microstructure and Liquidity*, 1(02), 1550008, the metaorder size distribution is observed directly, and the relation $\delta = \gamma$ also fails.

offer to the rest of the market, its flaws provide motivation for considering other approaches where the notion of a terminal price is absent, while keeping the idea of a martingale price.

16.2.1 Martingale Evolution of the Traded Price

One such alternative is the **Madhavan–Richardson–Roomans** (MRR) model. In contrast to the Glosten–Milgrom framework, the MRR model does not rely on the existence of a terminal time or a terminal price. Instead, it considers an underlying fundamental price $p_{\mathrm{F},t}$, which evolves over time to reflect both the information content of trades and unanticipated news arriving in the market. The setting of the MRR model is very close to that of the propagator model that we discussed in Chapter 13.

In the MRR model, $p_{\mathrm{F},t}$ coincides with the traded price, and its evolution consists of two terms: one corresponding to the information content of trades, and the other to a noise term ξ_t that captures unanticipated news:

$$p_{\mathrm{F},t} - p_{\mathrm{F},t-1} = G^* \times (\varepsilon_t - \widehat{\varepsilon_t}) + \xi_t, \tag{16.18}$$

where the information content of trades G^* is assumed to be constant (i.e. independent of time). The first term on the right-hand side of Equation (16.18) is indeed proportional to the **sign surprise** $\varepsilon_t - \widehat{\varepsilon_t}$, where (as in Section 13.3) $\widehat{\varepsilon_t}$ is defined as:

$$\widehat{\varepsilon_t} = \mathbb{E}_{t-1}[\varepsilon_t].$$

By construction, the fundamental price (and the traded price) $p_{\mathrm{F},t}$ is a martingale. Therefore, the long-term expectation of the fundamental price is always such that

$$\mathbb{E}_t[p_{\mathrm{F},t+T}] = p_{\mathrm{F},t}.$$

In this market, how should Bob set his quotes so as to have no ex-post regrets? Similarly to the Glosten–Milgrom model, the key to finding the solution for the MRR model is to note that for both a buy order and a sell order, the expected realised transaction price at $t+1$ should be equal to the fundamental price:

$$\begin{aligned} a_{t+1} &= \mathbb{E}_t[p_{\mathrm{F,t+1}}|\varepsilon_{t+1} = 1] = p_{\mathrm{F,t}} + G^*\left(1 - \widehat{\varepsilon}_{t+1}\right); \\ b_{t+1} &= \mathbb{E}_t[p_{\mathrm{F,t+1}}|\varepsilon_{t+1} = -1] = p_{\mathrm{F,t}} - G^*\left(1 + \widehat{\varepsilon}_{t+1}\right). \end{aligned} \tag{16.19}$$

Another way to think about these expressions is that other traders only trade with Bob when their information allows them to break even. This is of course a somewhat unrealistic assumption that can at best be true on average.

In the following sections, we work out several directly testable predictions of the MRR models, in particular concerning the response function $\mathcal{R}(\ell)$ considered throughout the book.

16.2.2 Correlation and Response in the MRR model

By Equation (16.19), it immediately follows that[8]

$$s_{t+1} = a_{t+1} - b_{t+1} = 2G^*,$$

which is, by assumption, constant in time. Similarly, the mid-price just before the $(t+1)^{\text{th}}$ trade is given by

$$m_{t+1} = \frac{a_{t+1} + b_{t+1}}{2} = p_{\text{F},t} - G^*\widehat{\varepsilon}_{t+1}.$$

Using Equation (16.18), the **mid-price dynamics** can be rewritten as

$$m_{t+1} - m_t = G^*\left(\varepsilon_t - \widehat{\varepsilon}_t\right) + G^*\left(\widehat{\varepsilon}_t - \widehat{\varepsilon}_{t+1}\right) + \xi_t. \tag{16.20}$$

This equation is somewhat similar to the propagator specification in Equa tion (13.22), but with an extra term on the right-hand side that is proportional to the change in the expected sign.

By considering a telescopic sum of expressions of the form in Equation (16.20), it follows that

$$m_{t+\ell} - m_t = G^* \sum_{n=t}^{t+\ell-1} \left(\varepsilon_n - \widehat{\varepsilon}_{n+1}\right) + \sum_{n=t}^{t+\ell-1} \xi_n. \tag{16.21}$$

By also using that $\mathbb{E}[\widehat{\varepsilon}_n \varepsilon_t] = \mathbb{E}[\varepsilon_n \varepsilon_t]$ when $n > t$, the **response function** $\mathcal{R}(\ell)$ (see Section 11.3.1) can then be expressed in terms of the sign correlation $C(\ell)$:

$$\mathcal{R}(\ell) = \mathbb{E}[(m_{t+\ell} - m_t)\cdot\varepsilon_t] = G^*(1 - C(\ell)) = \frac{s}{2}(1 - C(\ell)). \tag{16.22}$$

This remarkable relation can be tested empirically (as we will do in the next section).[9] When $\ell \to \infty$, the asymptotic value of $C(\ell)$ is 0. Therefore, the asymptotic value of the response function is

$$\mathcal{R}_\infty = G^* = s/2.$$

In the absence of any other costs, market-making ceases to be profitable when half the bid–ask spread only compensates the long-term impact of market orders. This is precisely the result we anticipated in Section 1.3.2.

Note that strictly speaking, the MRR model is a special case of the formalism that we have presented, with the additional assumption that the ε_t series is Markovian with $\widehat{\varepsilon}_t = \rho\varepsilon_{t-1}$, such that $C(\ell) = \rho^\ell$ with $\rho < 1$. (MRR did not consider the case where the ε_t series has long-range correlations).

[8] MRR consider the possibility that order processing costs c, or other costs incurred by the market-marker, can be added to the spread, to arrive at $s = 2(G^* + c)$. In this more general framework, the spread reflects both adverse impact and extra costs. We will neglect c throughout the rest of the chapter.

[9] In contrast with Equation (16.22), one finds that for the propagator model (for which the second term in Equation (16.20) is absent), the response function $\mathcal{R}(\ell)$ is a constant, equal to $1 - \mathbb{E}[\widehat{\varepsilon}^2]$. See Section 16.2.4.

16.2.3 Volatility versus Spread

In the MRR model, the long-term mid-price volatility per trade is given by

$$\widetilde{\sigma}^2_\infty = \lim_{\ell \to \infty} \frac{\mathbb{V}[m_{t+\ell} - m_t]}{\ell}, \\ = G^{*2}(1 - \mathbb{E}[\widehat{\varepsilon}^2]) + \Sigma^2,$$

where the tilde indicates that we work in trade time and Σ^2 is the variance of the noise term ξ_t in Equation (16.18). Approximating $\widehat{\varepsilon}_t$ by $C(1)\varepsilon_t$ and using the relation $s = 2G^*$, the above equation predicts an affine relation between the volatility per trade and the squared bid–ask spread:[10]

$$\widetilde{\sigma}^2_\infty = \frac{1 - C(1)^2}{4} s^2 + \Sigma^2. \tag{16.23}$$

We will perform empirical tests of this relation in Section 16.3. Observe that if φ denotes the number of trades per unit (calendar) time, then the relationship between the volatility per trade $\widetilde{\sigma}$ and the (usual) volatility per unit time σ is simply

$$\sigma = \widetilde{\sigma} \times \sqrt{\varphi}. \tag{16.24}$$

In the Markovian MRR case $C(\ell) = \rho^\ell$, an explicit formula can be derived for $\widetilde{\sigma}^2(\ell)$:

$$\widetilde{\sigma}^2(\ell) = \frac{\mathbb{V}[m_{t+\ell} - m_t]}{\ell} = G^{*2}\left[1 - \rho^2 - \frac{2\rho}{\ell}(1 - \rho^\ell)\right] + \Sigma^2. \tag{16.25}$$

Note in particular that $\widetilde{\sigma}^2_\infty \geq \widetilde{\sigma}^2(1)$ (i.e. short-term mid-price volatility does not exceed long-term volatility).

16.2.4 Interpolating Between the MRR and the Propagator Model

The MRR model surmises that sign surprises impact the *traded price*, whereas the propagator model assumes that sign surprises impact the *mid-price*. In order to illustrate the difference between these two models, it is conceptually useful to introduce a mixture model that interpolates between the propagator specification in Equation (13.19) and the MRR specification in Equation (16.20):

$$m_{t+1} - m_t = G^*(\varepsilon_t - \widehat{\varepsilon}_t) + G^{**}(\widehat{\varepsilon}_t - \widehat{\varepsilon}_{t+1}) + \xi_t, \tag{16.26}$$

where $G^{**} = G^*$ recovers the MRR model and $G^{**} = 0$ the propagator model. Alternatively, one can write this as an evolution for the traded price

$$p_t = m_t + G^* \varepsilon_t,$$

as

$$p_{t+1} - p_t = G^*(\varepsilon_{t+1} - \widehat{\varepsilon}_{t+1}) + (G^{**} - G^*)(\widehat{\varepsilon}_t - \widehat{\varepsilon}_{t+1}) + \xi_t, \tag{16.27}$$

where the second term vanishes in the MRR model, but leads to a “bid–ask bounce” (i.e. a short-term mean-reversion of the traded price) in the propagator model.

In this case, Equation (16.22) instead becomes:

$$\mathcal{R}(\ell) = G^*(1 - C(\ell)) + (G^{**} - G^*)(1 - \mathbb{E}[\widehat{\varepsilon}^2]), \tag{16.28}$$

which is an affine relation between $\mathcal{R}(\ell)$ and $1 - C(\ell)$.

[10] One can show that $\mathbb{E}[\widehat{\varepsilon}^2] = C(1)^2$ in fact holds with high numerical accuracy, with corrections smaller than 1%.

16.3 Empirical Analysis of the MRR Model

In this section, we perform empirical tests of two core predictions of the generalised MRR model: the relationship between the response function and the sign-correlation function (see Equation (16.22)) and the relationship between volatility and the bid–ask spread (see Equation (16.23)).

16.3.1 Response Function and Sign-Correlation Function

Figure 16.3 shows the values of $(1 - C(1)) \times \mathcal{R}(\ell)/\mathcal{R}(1)$ versus $(1 - C(\ell))$ for a large pool of stocks. The MRR model predicts the relationship between these quantities to be a straight line with slope 1 and intercept 0 (see Equation (16.22)). Overall, the MRR equation holds surprisingly well, but has some discrepancies:

- The MRR model appears to fare *better* for large-tick stocks than for small-tick stocks, in that the intercept of the regression is statistically different from zero (and weakly negative) and the slope is statistically larger than one for small-tick stocks. These biases can be explained by the mixture model; see Equation (16.28).
- A systematic positive (resp. negative) concavity appears for large ℓ for small-tick (resp. large-tick) stocks. This means that real price changes exhibit less (resp. more) resistance than what is apparent at small ℓ. As we discuss in the next chapter, this suggests that market-makers suffer more (resp. less) from adverse selection on medium time scales than on short time scales (see Equation (17.12)).

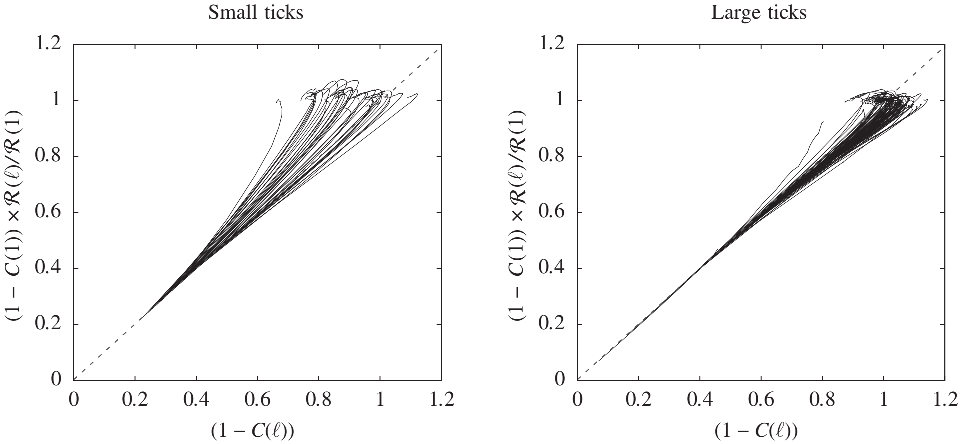

Figure 16.3. The quantity $(1 - C(1)) \times \mathcal{R}(\ell)/\mathcal{R}(1)$ versus $(1 - C(\ell))$ for a pool of 120 liquid stocks and $\ell \in \{1, 2, \cdots, 20\}$, divided into (left panel) small-tick stocks and (right panel) large-tick stocks. The dashed lines are regressions fitted to short lags only (such that $1 - C(\ell) \leq 0.6$), with slopes (left panel) 0.956 and (right panel) 1.01, and very small, statistically insignificant intercepts.

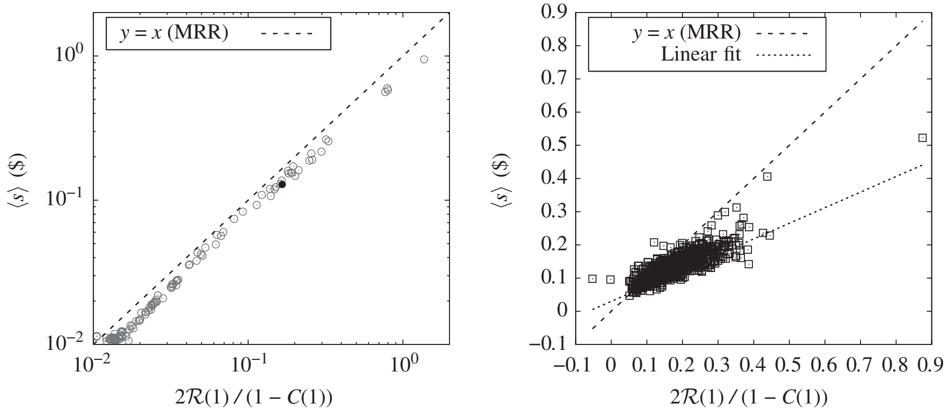

Figure 16.4. (Left panel) The average spread versus $2\mathcal{R}(1)/(1 - C(1))$ for a pool of 120 liquid stocks. Large-tick stocks are characterised by values of s close to \$0.01. Small-tick stocks are characterised by significantly larger values of the spread. The dashed line $y = x$ is the MRR prediction. (Right panel) The average spread $\langle s_t \rangle$ versus $2\mathcal{R}_t(1)/(1 - C_t(1))$, calculated for each 1-hour interval of TSLA trading during 2015. The dotted line is a linear fit $y = 0.47x + 0.03$. The dashed line is the MRR prediction $y = x$. The average point in the right plot corresponds to the black point in the left plot.

Note that while Figure 16.3 tests the functional relationship between $\mathcal{R}(\ell)$ and $1 - C(\ell)$, it does not question the MRR model's core prediction that $\mathcal{R}(1)$ should equal $s \times (1 - C(1))/2$. This is what we report in Figure 16.4, both cross-sectionally over our pool of stocks (left plot), and for one given small-tick stocks over time, for which the local spread varies appreciably (right plot). Overall, the MRR relation between spread, lag-1 impact and correlation is again remarkably well obeyed across different stocks (left plot). As we will argue in Chapter 17, this relationship is actually enforced by **competition between market-makers**, and states that the fastest market-makers break even on average.

The right panel in Figure 16.4 shows that although the MRR relation approximately holds *on average*, some significant deviations occur locally. On the one hand, market-makers do not increase spreads quite enough in situations where the impact of trades increases, or when the autocorrelation increases. On the other hand, the spread is systematically too wide in quieter situations, compensating the losses incurred in volatile situations.

16.3.2 Volatility and the Bid–Ask Spread

We now turn to the predicted affine relationship between long-term volatility and the bid–ask spread, Equation (16.23). We approximate the long-term volatility per trade $\widetilde{\sigma}_\infty$ by $\widetilde{\sigma}(\ell = 20)$, which is less noisy while still being close to $\widetilde{\sigma}_\infty$.

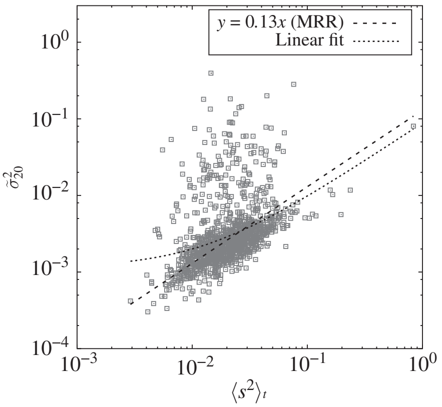

Figure 16.5. Scatter plot of the volatility $\widetilde{\sigma}_{20}^2$ versus average squared spread $\langle s^2 \rangle_t$ (measured in squared dollars) for one-hour intervals of TSLA, plotted in doubly logarithmic coordinates. The dash-dotted lines are linear regressions applied after smoothing the data with a kernel estimator, yielding $y = 0.087x + 0.0011$. The dotted line is the MRR prediction with $\Sigma^2 = 0$, leading to $y = 0.13x$.

Figure 16.5 shows a scatter plot of the squared volatility versus average squared spread $\langle s^2 \rangle_t$ for one-hour intervals during all trading days in 2015 for TSLA. The linear fit leads to $y = 0.087x + 0.0011$, which has a slope smaller than the slope 0.13 predicted by Equation (16.23) with $C(1) = 0.69$ (from Table 11.2). Note that with $\langle s^2 \rangle \approx 0.025$ for TSLA, one finds that the news component explains roughly a third of the price variance. This is similar to what Figure 13.2 (right) conveys in the context of the propagator model.

Interestingly, the plot suggests that there are two regimes: a relatively smooth cloud of points where the MRR relation is extremely well obeyed with a zero intercept (i.e. no contribution of the news component Σ^2) and more extreme volatility periods corresponding to scattered points, which are presumably dominated by exogenous events. These points outside the regular cloud are at the origin of the non-zero intercept of the linear fit, which agrees well with its intuitive interpretation.

Turning now to a cross-sectional test of Equation (16.23), we show in Figure 16.6 the average square volatility per trade $\widetilde{\sigma}_{20}^2$ as a function of $s^2 \times (1 - C(1)^2)$. The overall agreement with such a simple prediction is quite striking. In particular, the approximately linear relation between volatility per trade and spread is clearly vindicated.[11] The absence of any visible intercept suggests that the news contribution Σ^2 is itself proportional to the average squared spread. This means that

[11] This was first emphasised in Wyart, M., Bouchaud, J. P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, 8(1), 41–57.

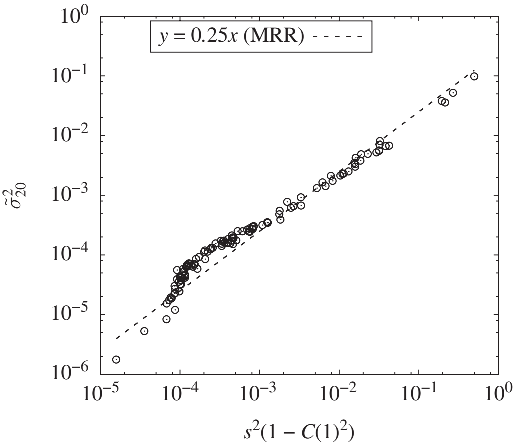

Figure 16.6. Cross-sectional scatter plot of $\widetilde{\sigma}_{20}^2$ versus $s^2 \times [1 - C(1)^2]$ (measured in squared dollars) for a pool of 120 liquid stocks traded on NASDAQ. Each point corresponds to one stock in 2015. The dashed line is the MRR prediction $y = x/4$.

all individual price movements (jumps included) are commensurate to the average spread.

16.4 Conclusion

We started this chapter with the traditional partitioning of traders into "market-makers" and "liquidity takers", a fraction of the latter possessing some information about the future value of the asset. In this context, the argument of Glosten and Milgrom states that the market-maker must open up a spread between the bid-price and the ask-price, in such a way that the losses incurred due to informed trades are compensated by spread gains (paid by the noise traders).

As we emphasised in the very first chapter of this book, markets can only function if liquidity providers are present. The Glosten–Milgrom model elicits a further constraint: liquidity providing is only viable if the fraction of informed trades is sufficiently small – or, equivalently, if the average amount of information in each trade is small. When the information gap is too large, market-makers retract, leading to a liquidity drought and a breakdown of the market. If the model is taken seriously, the fraction ϕ of trades that predict future price moves on a time horizon T should decrease at least as $1/\sqrt{T}$, with $\phi \approx 1\%$ for $T = 1$ day.

An important limitation of the model is that it assumes, as in the Kyle framework, the existence of a terminal time at which the fundamental price p_{F} (say the closing price of the market) is revealed and the game ends. In reality, some predictability exists over a wide spectrum of time scales, from high-frequency signals, which predict price moves over a few seconds, to intra-day signals and even signals on much longer horizons. The question of the profitability of market-making requires a framework where trading is open-ended, with no particular prediction horizon. We will discuss such a framework in the next chapter.

A simple setting where this question can be addressed is the Madhavan, Richardson and Roomans (MRR) model, which posits that the traded price p_t is a martingale, with returns linearly related to the sign of the order. This model is very close in spirit to the "surprise" version of the propagator model, which rather focuses on the mid-price m_t (see Section 13.3). The MRR makes falsifiable predictions about the impact response function $\mathcal{R}(\ell)$ and the volatility per trade $\widetilde{\sigma}$. In view of the simplicity of the model, the agreement with empirical data is remarkable.

In particular, spread and volatility appear as two sides of the same coin, with a causality that is hard to disentangle. Is spread merely compensating for volatility, as the Glosten–Milgrom arguments would have it, or is volatility also influenced by the spread, in the sense that the last transaction price shifts the expectations of the rest of the market? The success of the MRR model suggests that either a majority of trades in the markets are truly informed (which is doubtful), or that the market as a whole statistically adapts to the last traded price.

Take-Home Messages

(i) Marker makers are exposed to adverse selection because they offer to trade with other market participants, some of whom might be informed about the future value of the asset.

(ii) The Glosten–Milgrom framework is a stylised model for how market-makers should set their bid and ask quotes to exactly compensate for the costs of this adverse selection.

(iii) When liquidity takers are sufficiently uninformed (or misinformed), market-makers can solve the equations in the Glosten–Milgrom framework to find break-even quotes. In this framework, transactions always occur at the *a posteriori* fair price, and the price process is a martingale.

(iv) When liquidity takers are too informed, such a solution does not exist, so market-makers must withdraw from offering trades (or otherwise accept to make a loss on average).

(v) The MRR model is an example of a market where the Glosten–Milgrom conditions are met, and in which the transaction prices follow exactly a lag-1 propagator model.

(vi) The MRR model makes falsifiable predictions that agree very well with empirical data; in particular, the volatility per trade is found to be proportional to the spread.

16.5 Further Reading

The Determinants of the Bid–Ask Spread

Bagehot, W. (1971). The only game in town. *Financial Analysts Journal*, 27(2), 12–14.

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71–100.

Glosten, L. R., & Harris, L. E. (1988). Estimating the components of the bid/ask spread. *Journal of Financial Economics*, 21(1), 123–142.

Perold, A. F. (1988). The implementation shortfall: Paper versus reality. *The Journal of Portfolio Management*, 14(3), 4–9.

Subrahmanyam, A. (1991). Risk aversion, market liquidity, and price efficiency. *The Review of Financial Studies*, 4, 417–441

Krishnan, M. (1992). An equivalence between the Kyle (1985) and the Glosten-Milgrom (1985) models. *Economics Letters*, 40(3), 333–338.

Huang, R. D., & Stoll, H. R. (1997). The components of the bid–ask spread: A general approach. *Review of Financial Studies*, 10(4), 995–1034.

Handa, P., Schwartz, R., & Tiwari, A. (2003). Quote setting and price formation in an order driven market. *Journal of Financial Markets*, 6(4), 461–489.

Stoll, H. R. (2003). Market microstructure. In Constantinides, G. M., Harris, M., & Stulz, R. M. (Eds.), *Handbook of the economics of finance* (Vol. 1, pp. 553–604). Elsevier.

Amihud, Y., Mendelson, H., & Pedersen, L. H. (2006). Liquidity and asset prices. *Foundations and Trends' in Finance*, 1(4), 269–364.

Foucault, T., Pagano, M., & Röell, A. (2013). *Market liquidity: Theory, evidence, and policy*. Oxford University Press.

Tannous, G., Wang, J., & Wilson, C. (2013). The intra-day pattern of information asymmetry, spread, and depth: Evidence from the NYSE. *International Review of Finance*, 13(2), 215–240.

Break-Even Conditions and Metaorder Impact

Donier, J. (2012). Market impact with autocorrelated order flow under perfect competition. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2191660.

Farmer, J. D., Gerig, A., Lillo, F., & Waelbroeck, H. (2013). How efficiency shapes market impact. *Quantitative Finance*, 13(11), 1743–1758.

see also Rogers, K., https://mechanicalmarkets.wordpress.com/2016/08/15/price-impact-in-efficient-markets/.

The MRR Model and Extensions

Madhavan, A., Richardson, M., & Roomans, M. (1997). Why do security prices change? A transaction-level analysis of NYSE stocks. *Review of Financial Studies*, 10(4), 1035–1064.

Wyart, M., Bouchaud, J. P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008). Relation between bid-ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, 8(1), 41–57.

Bonart, J., & Lillo, F. (2016). A continuous and efficient fundamental price on the discrete order book grid. https://ssrn.com/abstract=2817279.

The Role of the Tick Size

Harris, L. E. (1994). Minimum price variations, discrete bid-ask spreads, and quotation sizes. *Review of Financial Studies*, 7(1), 149–178.

Bessembinder, H. (2000). Tick size, spreads, and liquidity: An analysis of NASDAQ securities trading near ten dollars. *Journal of Financial Intermediation*, 9(3), 213–239.

Goldstein, M. A., & Kavajecz, K. A. (2000). Eighths, sixteenths, and market depth: Changes in tick size and liquidity provision on the NYSE. *Journal of Financial Economics*, 56(1), 125–149.

Zhao, X., & Chung, K. H. (2006). Decimal pricing and information-based trading: Tick size and informational efficiency of asset price. *Journal of Business Finance & Accounting*, 33(5–6), 753–766.

Dayri, K., & Rosenbaum, M. (2015). Large tick assets: Implicit spread and optimal tick size. *Market Microstructure and Liquidity*, 1(01), 1550003.

Bonart, J. (2016). What is the optimal tick size? A cross-sectional analysis of execution costs on NASDAQ. https://ssrn.com/abstract=2869883.