17

The Profitability of Market-Making

A market-maker knows the price of everything and the value of nothing.

(After Oscar Wilde)

In Chapter 16, we discussed two models in which the transaction price is a martingale. This encodes in a strict manner that both the market-maker and the other trader have on average “no ex-post regrets” about their trades, because the price of each trade is also the best estimate of future prices. In this set-up, there is no arbitrage and no profits for market-makers. Although extremely convenient, one might ask whether this theoretical framework is too strict to account for the behaviour that occurs in real markets.

In this chapter, we revisit the question of whether simple market-making strategies can be profitable. Throughout our discussion, we employ as little theoretical prejudice as possible. This allows us to discuss issues such as inventory constraints and holding times, which are absent from the martingale approach.

Before embarking into this agnostic, model-free analysis, we first rephrase the **martingale** hypothesis for traded prices in a more illuminating way, which makes it obvious that market-making cannot, in this case, be profitable. The basic observation is as follows. Since the conditional expectation of *all* future traded prices is equal to the last traded price p_t, this is true in particular for the expectation of the next traded price at the bid and the next traded price at the ask. More precisely, let us denote by $t_a > t$ (respectively, $t_b > t$) the time of the next trade at the ask (respectively, bid), and $a_{\text{next}} := a_{t_a}$ the corresponding value of the ask-price (respectively, $b_{\text{next}} := b_{t_b}$ the corresponding value of the bid-price). Then:

$$\mathbb{E}[a_{\text{next}}] = \mathbb{E}[b_{\text{next}}] = p_t.$$

Therefore, although the next trade may happen either at the bid or at the ask, these two prices must both on average equal p_t, which is a surprising result given that $b_t < a_t$ at all times!

The intuition behind this apparent puzzle is the following: if the next market order is a buy, then the corresponding ask-price is most likely to be larger than p_t. If instead one or more sell market orders arrive first, they will impact the price downwards, such that the next buy order will hit a depressed ask-price, $a_{\text{next}} \leq a_t$. On average, the two effects cancel out in a competitive setting and $\mathbb{E}[a_{\text{next}}] = p_t = \mathbb{E}[b_{\text{next}}]$, as we now demonstrate.[1]

Let us assume that $\mathbb{E}[a_{\text{next}}] > \mathbb{E}[b_{\text{next}}]$. In this case, one could devise a market-making strategy that would profit from the difference: imagine observing the best bid and ask quotes posted by other traders, then simultaneously pegging infinitesimal (impact-less) buy and sell quantities at the bid- and ask-prices, respectively, and waiting for execution. The pegged order to buy (respectively, sell) will by definition be executed at b_{next} (respectively, a_{next}), so the expected profit of the round-trip is $\mathbb{E}[a_{\text{next}} - b_{\text{next}}]$. If this quantity was positive on average, it would mean that the market-makers are not competitive enough, because they could offer better quotes; if it was negative on average, market-makers would lose money on average and would have to provide less aggressive quotes. Hence, necessarily $\mathbb{E}[a_{\text{next}}] = \mathbb{E}[b_{\text{next}}]$.

Note that the above argument implicitly assumes that any limit order placed at the bid-price or at the ask-price is executed against the next market order. We will use this hypothesis repeatedly in the next section. The hypothesis is reasonable for small-tick stocks, where queue-priority effects are negligible. This can be understood in two different ways. The first is that for small-tick stocks, the queue lengths at the best quotes are usually short. Because the probability that an arriving market order consumes the whole best queue is large, traders who place limit orders at the best quotes are very likely to receive a matching, even if their limit orders are at the back of the queue. The second is that if the relative tick size is small, it costs a trader very little to place an order one tick ahead of the prevailing best bid- or ask-prices, thereby effectively jumping the queue (although this explanation neglects the possibility that other participants can undercut this new limit order by placing other limit orders at even better prices).

For large-tick stocks, on the other hand, queues are long and priority considerations become crucial. Both of the arguments about small-tick stocks fail: placing a limit order at the best quotes typically means being last in a long queue, with a very small probability of executing against the next market order, and it is usually impossible to jump the queue because there is typically no gap between the bid and the ask. For large-tick stocks, market-makers must therefore compete for priority, rather than competing for spreads.

[1] This argument is originally due to Jaisson, T. (2015). Liquidity and impacts in fair markets. *Market Microstructure and Liquidity*, 1(02), 1550010.

17.1 An Infinitesimal Market-Maker

In line with the argument of the previous section, imagine a market-maker that submits limit orders to maintain a very small volume υ_0 at both the bid- and ask-prices. For now, we assume that the market-maker has no inventory constraints, and can therefore conduct any number of buy or sell trades consecutively.

What is the **Profit & Loss** (P&L) balance of this market-maker? The market-maker's t^{th} transaction takes place either at a_t (when the sign of the market order is $\varepsilon_t = +1$) or at b_t (when the sign of the market order is $\varepsilon_t = -1$). One way to assign a value to each of the market-maker's transactions is to compare their price to the mid-price at some later time T. In this framework, the gain or loss from the t^{th} individual trade, **marked-to-market** at time T, can be written as

$$g_t = \upsilon_0 \varepsilon_t \left(m_t + \varepsilon_t \frac{s_t}{2} - m_T \right). \tag{17.1}$$

To understand Equation (17.1), consider a buy market order ($\varepsilon_t = +1$), which hits the ask-queue at time t. When this trade occurs, the money received by the market-maker is $\upsilon_0 a_t = \upsilon_0 \times (m_t + s_t/2)$. The market-maker is then short υ_0 of the asset. At some later time T, the value of this position is marked-to-market at the mid-price m_T, at value $-\upsilon_0 m_T$.[2] The same arguments apply for a sell market order.

We will also assume that there is an additional payment ϖ associated with each trade. This payment could be a processing or transaction cost (for which $\varpi < 0$), a rebate (for which $\varpi > 0$), or could simply be zero. The effect of this additional payment is to create an extra gain or loss for the market-maker, equal to $\upsilon_0 \varpi$.

In this framework, the total P&L of the market-maker, marked to the mid-price at time T, is given by

$$\mathcal{G}_T = \upsilon_0 \left[\sum_{t=0}^{T-1} \theta_t \varepsilon_t \left(m_t + \varepsilon_t \frac{s_t}{2} - m_T \right) + \theta_t \varpi \right], \tag{17.2}$$

where θ_t is an **execution indicator**

$$\theta_t = \begin{cases} 1, & \text{if the limit order was matched at } t, \\ 0, & \text{otherwise.} \end{cases}$$

There are no partial fills when $\upsilon_0 = 1$ lot. Therefore, upon a market order arrival, the market-maker's limit order at the relevant best quote either fully matches (to the incoming market order) or does not match at all (due to its unfavourable position in the queue). For very small-tick stocks, the size of the queue at the best quotes is typically very small, so it is reasonable to assume that $\theta_t = 1$ most of the time. For large-tick stocks, it is much more difficult to obtain a position near the front of

[2] This assumption is not crucial in the following, as it would only change a boundary term to the total P&L of the market-maker, which is negligible when the market-maker inventory is bounded.

the queue, and large market order arrivals are also more rare. For these large-tick stocks, it is more reasonable to assume that $\theta_t = 0$ most of the time, except when one manages to be at the front of the queue.

By taking expectations of Equation (17.2), we arrive at the expected gain of our market-maker between times 0 and T:

$$\mathbb{E}[\mathcal{G}_T] = \upsilon_0 \left(\sum_{t=0}^{T-1} \mathbb{E}\left[\theta_t \varepsilon_t \left(m_t + \varepsilon_t \frac{s_t}{2} - m_T \right) + \theta_t \varpi \right] \right).$$

To proceed with our calculations, we now assume that θ_t is uncorrelated with the value of the spread and the price history.[3] Within this approximation,

$$\mathbb{E}[\mathcal{G}_T] = \upsilon_0 \mathbb{E}[\theta] \sum_{t=0}^{T-1} \left(\frac{\mathbb{E}[s]}{2} + \varpi - \mathcal{R}(T-t) \right),$$

where we have used the definition of the response function, Equation (16.22).

From this formula, it is clear that when $\varpi > 0$ (i.e. when the liquidity taker pays an additional fee per trade), the effect is tantamount to an upward shift of the ask-price and downward shift of the bid-price, both of size $\varpi > 0$, leading to an increased effective spread from s to $s + 2\varpi$.

Assuming that $\mathcal{R}_\infty$ exists and is finite, one has, for very large T:

$$\sum_{t=0}^{T-1} \mathcal{R}(T-t) = \sum_{u=1}^{T} \mathcal{R}(u) \approx T\mathcal{R}_\infty.$$

The expected long-term gain of the market-maker is thus given by a balance between spread, fees and long-term impact:

$$\mathbb{E}[\mathcal{G}_T] \approx T\upsilon_0 \mathbb{E}[\theta] \left(\frac{\mathbb{E}[s]}{2} + \varpi - \mathcal{R}_\infty \right). \tag{17.3}$$

Recalling the results of Section 13.2, we have seen that when $\beta > (1-\gamma)/2$, impact decays so quickly that the price is sub-diffusive (i.e. mean-reverting) and $\mathcal{R}(\ell)$ diverges to $-\infty$ in the limit $\ell \to \infty$. This situation would be a boon for market-makers, since Equation (17.3) predicts a strongly positive gain in that case. If, however, it instead holds that $\beta < (1-\gamma)/2$, then impact decay is too slow to prevent price trends. In this case, $\mathcal{R}(\ell)$ instead diverges to $+\infty$ when $\ell \to \infty$, rendering market-making extremely difficult. The important take-home message is that market-making is easy when prices mean-revert but difficult when prices trend. We will present a similar discussion when comparing the profitability of market orders to that of limit orders in Section 21.3.

[3] This assumption is reasonable for small-tick assets, for which $\theta_t \approx 1$, but is not justified for large-tick assets, for which the execution probability increases either when a large market order arrives or when the queue is very short. This leads to increased adverse selection, which we neglect here – but see Section 17.3.

17.2 Inventory Control for Small-Tick Stocks

An important problem with the result in Equation (17.3) is that it completely disregards the fact that the market-maker's inventory imbalance can become arbitrarily large. Even if market order signs were uncorrelated, the random imbalance between buy and sell volumes up to time T would grow like $\sqrt{T}$, which would lead to an unbounded market-maker inventory at large times. In real markets, market order signs are long-range autocorrelated, with an autocorrelation function $C(\ell) \sim l^{-\gamma}$, with $\gamma < 1$, which leads to an imbalance that grows even more quickly, as $T^{(1-\gamma)/2} \gg \sqrt{T}$. In any case, this imbalance diverges, positively or negatively, when $T \to \infty$. The risk of the market-maker therefore also diverges, which indicates that market-makers face the possibility of ruin due to an adverse market move. To account for this risk, it is necessary for market-makers to implement some form of **inventory control**. In this section, we consider the question of calculating a market-maker's P&L when applying such inventory-control limits.

17.2.1 The Problem

Deriving the optimal strategy of our market-maker Bob in the presence of inventory risk and long-range autocorrelations of market order signs is a very difficult problem. To gain some intuition, we will assume that the Bob's inventory-risk control amounts to him following a mean-reversion strategy that consists of offering less liquidity at the ask and more liquidity at the bid if he is already short, and vice-versa if he is already long.

Let ψ_{t-1} denote Bob's net position after trade $t-1$. Given ψ_{t-1}, we assume that the market-maker participates with a volume $\upsilon_t(\varepsilon)$ at the ask ($\varepsilon = +1$) and at the bid ($\varepsilon = -1$), with:

$$\upsilon_t(\varepsilon) = \upsilon_0 \max(1 + \kappa\psi_{t-1}\varepsilon, 0), \tag{17.4}$$

where $\kappa > 0$ is a parameter that describes how tightly Bob controls his inventory. To understand Equation (17.4), imagine that Bob's inventory is $\psi_{t-1} = 1$, (i.e. he is net long one unit). At time t, he will participate with volume proportional to $1+\kappa$ at the ask and with volume proportional to $1-\kappa$ at the bid, such that he is more likely to sell and reduce his inventory than to buy and increase his inventory. An alternative inventory control strategy is described in Appendix A.5.

Recall that for a small-tick stock, one can assume that $\theta_t = 1$ for most t. In this situation, the evolution equation for ψ_t becomes

$$\begin{aligned}\psi_t - \psi_{t-1} &= -\upsilon_t(\varepsilon_t)\theta_t\varepsilon_t,\\ &\approx -\upsilon_t(\varepsilon_t)\varepsilon_t.\end{aligned}$$

Mathematically, the difficulty with Equation (17.4) is the truncation at zero, which prevents the market-maker from offering negative volumes (which would have no meaning). To proceed with analytical calculations, we simply drop that constraint,[4] which amounts to neglecting the possibility that $\kappa|\psi|$ exceeds 1. This simplification allows one to write the evolution of the inventory ψ as a discrete **Ornstein–Uhlenbeck** (or autoregressive) process:

$$\psi_t = (1 - \kappa\upsilon_0)\psi_{t-1} - \upsilon_0\varepsilon_t. \tag{17.5}$$

Introducing $\alpha := 1 - \kappa\upsilon_0$ yields the solution

$$\psi_t = -\upsilon_0 \sum_{\ell=0}^{\infty} \alpha^{\ell} \varepsilon_{t-\ell}. \tag{17.6}$$

Since we assume that the mean of the order signs is 0, the mean inventory is also 0. Its unconditional variance is given by

$$\mathbb{V}[\psi] = \upsilon_0^2 \frac{1 + 2\sum_{\ell=1}^{\infty} \alpha^{\ell} C(\ell)}{1 - \alpha^2}, \tag{17.7}$$

where $C(\ell)$ is the market order sign autocorrelation function. The expression (17.6) shows that in the presence of inventory control, the position of the market-maker remains finite, and only diverges when $\kappa \to 0$ (i.e. when $\alpha \to 1$).

Neglecting the fees for now (i.e. assuming that $\varpi = 0$) and assuming zero initial position (i.e. setting $\psi_0 = 0$), the total expected P&L of the market-maker can be expressed as

$$\begin{aligned}
\mathbb{E}[\mathcal{G}] &= \psi_T m_T - \sum_{t=0}^{T-1} \mathbb{E}\left[(\psi_t - \psi_{t-1}) \cdot \left(m_t + \frac{\varepsilon_t s_t}{2}\right)\right] \\
&\approx T\mathbb{E}\left[\left(\upsilon_0\varepsilon_t - \kappa\upsilon_0^2 \sum_{\ell=0}^{\infty} \alpha^{\ell} \varepsilon_{t-1-\ell}\right) \cdot \left(m_t + \frac{\varepsilon_t s_t}{2}\right)\right],
\end{aligned}$$

where we have neglected in the second equation the sub-leading boundary term corresponding to the value of the inventory at $t = T$. This is justified because the inventory remains finite as long as $\kappa > 0$ (i.e. the inventory does not grow with T, see Equation (17.7)).

Let us first look at the term coming from the spread. Neglecting any possible correlations between the spread and market order signs, one finds

$$\frac{\mathbb{E}[s]}{2}\upsilon_0\mathbb{E}\left[\left(1 - \kappa\upsilon_0 \sum_{\ell=0}^{\infty} \alpha^{\ell} \varepsilon_{t-1-\ell}\varepsilon_t\right)\right] = \frac{1-\alpha}{2\alpha}\mathbb{E}[s]\upsilon_0 \sum_{\ell=1}^{\infty} \alpha^{\ell}(1 - C(\ell)). \tag{17.8}$$

[4] We will return to our discussion of this approximation at the end of this section.

The impact term is slightly more subtle; it reads:

$$\upsilon_0 \mathbb{E}\left[\left(\varepsilon_t - \kappa\upsilon_0 \sum_{\ell=0}^{\infty} \alpha^{\ell} \varepsilon_{t-1-\ell}\right) \cdot m_t\right] = \upsilon_0 \mathbb{E}\left[\varepsilon_t \cdot m_t\right] - \upsilon_0 \frac{1-\alpha}{\alpha} \left\langle \sum_{\ell=1}^{\infty} \alpha^{\ell} \mathbb{E}\left[\varepsilon_{t-\ell} \cdot m_t\right] \right\rangle. \tag{17.9}$$

In the second term on the right-hand side, we write as an identity:

$$m_t \equiv m_t - m_{t-\ell} + m_{t-\ell},$$

and recall that by definition

$$\mathcal{R}(\ell) = \mathbb{E}\left[\varepsilon_{t-\ell} \cdot (m_t - m_{t-\ell})\right]. \tag{17.10}$$

Therefore, one can further transform the above expression as:

$$\upsilon_0 \mathbb{E}\left[\left(\varepsilon_t - \kappa\upsilon_0 \sum_{\ell=0}^{\infty} \alpha^{\ell} \varepsilon_{t-1-\ell}\right) \cdot m_t\right] = -\upsilon_0 \frac{1-\alpha}{\alpha} \sum_{\ell=1}^{\infty} \alpha^{\ell} \mathcal{R}(\ell), \tag{17.11}$$

where we have used the fact that $\mathbb{E}\left[\varepsilon_t \cdot m_t\right] = \mathbb{E}\left[\varepsilon_{t-\ell} \cdot m_{t-\ell}\right]$.

Gathering all the terms, we finally recover the profit per unit transaction and unit volume of an inventory-constrained market-maker:

$$\frac{\mathbb{E}[\mathcal{G}]}{\upsilon_0 T} \approx \frac{1-\alpha}{\alpha} \left(\frac{1}{2} \mathbb{E}[s] \sum_{\ell=1}^{\infty} \alpha^{\ell} (1 - C(\ell)) - \sum_{\ell=1}^{\infty} \alpha^{\ell} \mathcal{R}(\ell) \right). \tag{17.12}$$

This expression is only approximate, because we have neglected the non-linearity in Equation (17.4). Besides, the mean-reversion strategy (17.4) has no reason to be optimal, so more profitable strategies could possibly exist. Still, in spite of all our approximations, we recover that for the benchmark MRR case, for which $\mathcal{R}(\ell) = (1 - C(\ell))\mathbb{E}[s]/2$, the average profit of the market-maker is equal to 0 for any value of α. This makes sense in this case: since the transaction price is fair, any market-making strategy must break even.

We briefly return to the validity of our approximation that relies on the assumption that $|\kappa\psi| \ll 1$ with large probability, which we rephrase as $\kappa^2 \mathbb{V}(\psi) \ll 1$. In the limit $\kappa \to 0$, and assuming that $C(\ell) \sim c_\infty/\ell^\gamma$ for large ℓ, with $0 < \gamma < 1$, one finds

$$\kappa^2 \mathbb{V}[\psi] = (\kappa\upsilon_0)^2 \frac{1 + 2\sum_{k=1}^{\infty} \alpha^k C(k)}{1-\alpha^2} \approx c_\infty \Gamma[1-\gamma](\kappa\upsilon_0)^\gamma \to 0.$$

Therefore, our approximation that $1 - \kappa\psi$ is positive is clearly supported for slow market-making.

In the other limit $\alpha = 1 - \kappa\upsilon_0 \to 0$, one finds:

$$\kappa^2 \mathbb{V}[\psi] = (\kappa\upsilon_0)^2 \frac{1 + 2\sum_{k=1}^{\infty} \alpha^k C(k)}{1-\alpha^2} \approx 1,$$

so our approximation is questionable here. However, in this limit, Equation (17.5) simply reads

$$\psi_t = -\upsilon_0 \varepsilon_t,$$

so that

$$\kappa|\psi| = \upsilon_0\kappa \to 1, \text{ with } \upsilon_0\kappa < 1.$$

Therefore, the approximation remains true in this limit as well, and recovers the exact criterion in the MRR case. In the intermediate cases however, the hypothesis has no reason to hold precisely.

17.2.2 Break-Even Conditions and Empirical Results

We now study two extreme limits of the general formula in Equation (17.12). We consider both expressions with ϖ set to zero; for $\varpi \neq 0$ one simply needs to replace s by $s + 2\varpi$.

(i) **Slow market-making** corresponds to a weakly mean-reverting inventory (i.e. $\kappa \to 0$ and thus $\alpha \to 1$). In this limit, and assuming that $\mathcal{R}_\infty$ is finite, one recovers to leading order the result of the last section:

$$\frac{\mathbb{E}[\mathcal{G}]}{\upsilon_0 T} = \frac{1}{2}\mathbb{E}[s] - \mathcal{R}_\infty. \tag{17.13}$$

This expression is the negative of the average ex-post gain of a market order, which pays the half-spread but makes the long-term average impact $\mathcal{R}_\infty$. In the MRR model, $\mathcal{R}_\infty = s/2$ so the slow market-maker makes no profit.

(ii) **Fast market-making** corresponds to a strongly mean-reverting inventory (i.e. $\kappa\upsilon_0 \to 1$ and thus $\alpha \to 0$), where the inventory only depends on the last trade. In this limit

$$\frac{\mathbb{E}[\mathcal{G}]}{\upsilon_0 T} = \frac{1}{2}\mathbb{E}[s](1 - C(1)) - \mathcal{R}(1). \tag{17.14}$$

Within this framework, the least risky market-making strategy (which corresponds to the fastest possible mean-reversion to zero inventory) has zero average gain provided

$$\mathbb{E}[s] = \frac{2\mathcal{R}(1)}{1 - C(1)}, \tag{17.15}$$

which is precisely the MRR **break-even condition** (see Equation (16.22)) for $\ell = 1$. Hence, as noted in the last section, the MRR relation is such that market-makers break even on average on all time scales. For $\ell = 1$, the intuition here is that when trades tend to be strongly correlated ($C(1) \to 1$), Bob cannot easily unwind his position, and the adverse impact is much larger than the short-term impact $\mathcal{R}(1)$ (see also the discussion in Appendix A.5).

Figure 17.1 shows the empirically determined P&L of Bob's simple strategy normalised by the average spread, as a function of the inventory control parameter $\alpha = 1 - \kappa\upsilon_0$, where Equation (17.12) is computed using the empirically determined $C(\ell)$ and $\mathcal{R}(\ell)$. If one does not account for rebates, the average gain of Bob's simple strategy is always negative, with fast market-making strategies ($\alpha = 0$) faring better

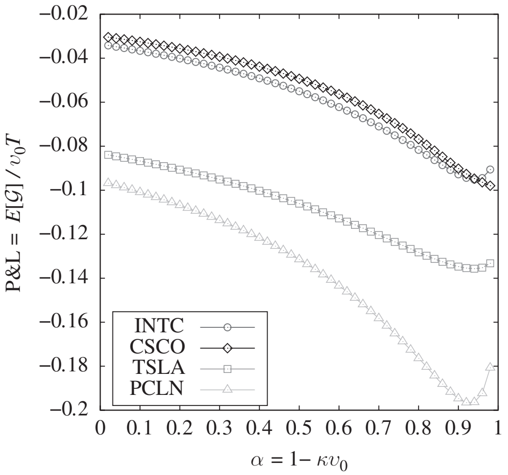

Figure 17.1. The normalised P&L from Equation (17.12), divided by the average spread, as a function of the inventory control parameter $\alpha = 1 - \kappa\upsilon_0$, and with the rebate fees ϖ set to zero, for INTC, CSCO, TSLA and PCLN. Note that in the last two cases, which correspond to large-tick stocks, Equation (17.12) is a priori not warranted (see Section 17.3).

than slow market-making strategies. The cost per trade is of the order of $0.10\langle s\rangle$ for the small-tick stocks, which is typically larger than the rebate fee $\varpi \approx \$0.0025$. This suggests that limit orders are in fact used as a cheaper alternative to market orders for small-tick assets, leading to spreads that are too small on average. For large-tick stocks, a more detailed discussion is needed (see Section 17.3).

Figure 17.2 shows the phase diagram of market-making in the plane $\mathcal{R}(1)$ vs. $\mathbb{E}[s]/2$, with two lines of slope $\mathcal{R}_\infty/\mathcal{R}(1)$ and $(1 - C(1))^{-1}$, corresponding to profitable slow market-making, and profitable fast market-making, respectively. In the MRR model, the two slopes are equal.

17.3 Large-Tick Stocks

For large-tick stocks, the spread is usually equal to its minimum value of one tick ϑ – which is, by definition, large. Naively, this should benefit market-makers, since the spread is bounded from below and is therefore artificially large. However, Bob's life is not that easy. His P&L (without inventory control) can be written as

$$\mathbb{E}[\mathcal{G}_T] = \upsilon_0 T\left[\mathbb{E}[\theta_t]\frac{\vartheta}{2} - \sum_{t=0}^{T-1}\mathbb{E}[\theta_t\varepsilon_t(m_T - m_t)]\right].$$

The position of Bob in the queue is now crucial. If he manages to always be at the front of the queue (for example if he is the only market-maker in town), he will

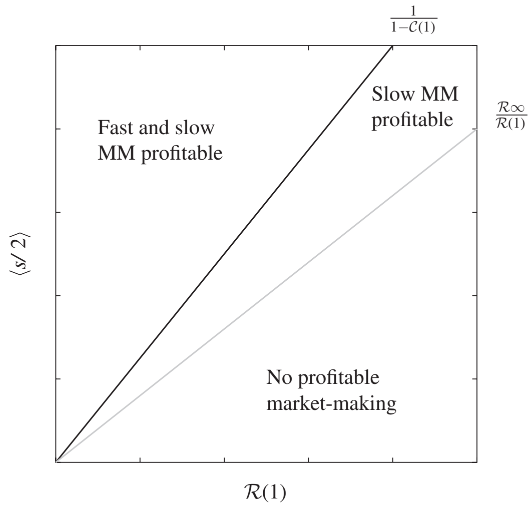

Figure 17.2. The theoretical bounds on the spread as a function of $\mathcal{R}(1)$. The grey line shows the slow market-making profitability bound $\mathcal{R}(\infty)/\mathcal{R}(1)$: slow market-making is profitable *above* this line. The black line shows the fast market-making bound slope $1/(1-C(1))$: fast market-making becomes profitable *above* this line. Note that some markets are such that the black line is below the grey line.

be executed against *all* incoming market orders, so $\theta_t = 1$. In this case, we recover exactly the same result as (17.3) for large T,

$$\mathbb{E}[\mathcal{G}_T^{\text{top}}] = \upsilon_0 T \left[\frac{\vartheta}{2} - \mathcal{R}_\infty\right],$$

for **top-priority limit orders**. In the presence of inventory control, Equation (17.12) also continues to hold if Bob is always at the front of the queue. Hence, the MRR relation $\mathcal{R}(\ell) = (1 - C(\ell))\vartheta/2$ ensures that even the smartest market-maker, always at the front of the queue, breaks even (in the absence of further rebate fees).

Now, consider that Bob has a random position in the queue. His limit order will only be executed at the next transaction if the incoming market order is large enough, but in these cases one expects the impact of these market orders to be larger than the unconditional impact across all market orders. More precisely, it is interesting to study the conditional response function

$$\mathcal{R}_\phi(\ell) = \langle \varepsilon_t \cdot (m_{t+\ell} - m_t) | \upsilon_t \geq \phi V_t \rangle, \tag{17.16}$$

where υ_t is the volume of the incoming market order and V_t is the total volume at the best quote. In words, this response function computes the impact of a market order that consumes more than a fraction ϕ of the available volume. In particular, one recovers the usual unconditional response function when $\phi = 0$ and the response function of price-changing market orders when $\phi = 1$:

$$\mathcal{R}_{\phi=0}(\ell) \equiv \mathcal{R}(\ell), \qquad \mathcal{R}_{\phi=1}(\ell) \equiv \mathcal{R}^1(\ell).$$

In Figure 17.3, we show $\mathcal{R}_\phi(\ell)$ as a function of ϕ for $\ell = 1$ (corresponding to $\alpha = 0$) and $\ell = 20$ (corresponding to $\alpha = 0.95$), for our two large-tick stocks. One clearly observes that $\mathcal{R}_\phi(\ell)$ is an increasing function of ϕ, which confirms that impact is larger for more aggressive market orders. Now, consider a limit order with a relative queue position equal to ϕ. The corresponding execution indicator θ_t is equal to 1 when $\upsilon_t \geq \phi V_t$ and equal to 0 otherwise. Since $\mathcal{R}_\phi(\ell)$ is an increasing function of ϕ, the following inequality holds true:

$$\mathbb{E}[\theta_t]\mathbb{E}\left[\varepsilon_t(m_T - m_t)\right]] \leq \mathbb{E}\left[\theta_t\varepsilon_t(m_T - m_t)\right]],$$

where the equality holds only for top-priority limit orders. The adverse selection suffered by a randomly placed limit order is stronger than the adverse selection of a top priority limit order, because there is a positive correlation between the probability to be executed and a large subsequent adverse price move. Hence:

$$\mathbb{E}[\mathcal{G}_T^{\text{random}}] < \mathbb{E}[\theta] \times \mathbb{E}[\mathcal{G}_T^{\text{top}}].$$

Looking again at Figure 17.1, one observes that for large-tick stocks, and for a fast market-maker ($\alpha \to 0$) with top priority,

$$\frac{\mathbb{E}[\mathcal{G}_T^{\text{top}}]}{\upsilon_0 T} \approx -0.05\vartheta.$$

In the NASDAQ markets,[5] typical rebate fees are of the order of 0.25ϑ, leaving an average total gain per trade equal to about 0.2ϑ for top priority limit orders on large-tick stocks.[6] As the priority of the limit order degrades, one expects adverse selection costs to increase until the gain per trade (including rebate fees) vanishes. From Figure 17.3, one estimates a difference of about 0.2ϑ between the average of $\mathcal{R}_\phi(20)$ over all values of ϕ (corresponding to a randomly placed limit order) and the unconditional response $\mathcal{R}(20) = \mathcal{R}_0(20)$ (corresponding to a top priority limit order). Hence, randomly placed limit orders roughly break even on average.

17.4 Conclusion

Market-makers earn the spread but suffer from price impact. By formalising the basic Glosten–Milgrom intuition in a model-free framework, we have shown how

[5] Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *Review of Financial Studies*, 27(8), 2267–2306.

[6] This order of magnitude will be confirmed by an independent calculation in Section 21.4.

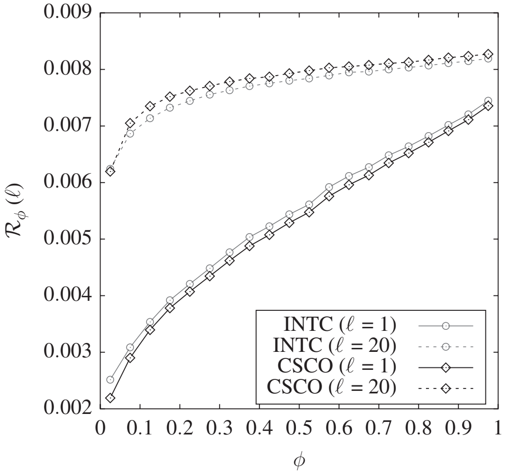

Figure 17.3. The response function $\mathcal{R}_\phi(\ell) = \langle \varepsilon_t \cdot (m_{t+\ell} - m_t) | \upsilon_t > \phi V_t \rangle$ as a function of ϕ for (circles) INTC and (diamonds) CSCO. The solid curves show the results for $\ell = 1$ and the dotted curves show the results for $\ell = 20$.

to compute the long-term P&L of a market-maker who is always at the best quotes and manages inventory risk. The market-maker's gains depend on the relative size of the spread and on the lag-dependent response function. For loosely controlled inventory risk, corresponding to slow market-making, the balance is between the half-spread and the long-term impact. For tightly controlled inventories, the balance is between the half-spread and the short-term impact, corrected by the autocorrelation of the trades. Within the MRR model, spread and impact turn out to balance each other exactly, for any market-making frequency. Reality is more complex, however.

Empirical data confirms that market-making in modern electronic markets is highly competitive. Simple market-making strategies with inventory control are found to yield negative profits (see Figure 17.1). Smarter market-making strategies, which include high-frequency signals that help market-makers to decide more precisely when and where to submit or cancel a limit order, can presumably be made to eke out a small profit. For large-tick stocks, the main challenge is to gain time priority in the queue: top-priority limit orders are found to be profitable on average, after including rebate fees.

In any case, the main practical conclusion is that, like in the MRR model, market orders and limit orders incur on average roughly equivalent costs – at least in the absence of short-term information about the order flow and price changes (see Section 21.3). This equivalence can be seen as a consequence of competition between market-makers, which compresses the spread to its minimum value.

Take-Home Messages

(i) If prices are martingales, the expected price of the next trade that will take place at the ask must equal the expected price of the next trade that will take price at the bid.

(ii) This illustrates that market-making is a more subtle task than it may first appear. The main difficulty of market-making stems from the problem that market-makers cannot choose the execution time and price at which they trade – and thus earn the instantaneous bid–ask spread.

(iii) Price impact and fees both offset the potential profits that market-makers could earn.

(iv) For large-tick stocks, trade prices are not a martingale, due to the frequent occurrence of bid–ask bounce. However, obtaining limit order executions for these stocks is difficult, because it typically requires waiting in a long queue. Therefore, the profitability of market-making for such stocks is not obvious.

17.5 Further Reading

Trading Costs and the Profitability of Market-Making

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.

Handa, P., Schwartz, R. A., & Tiwari, A. (1998). The ecology of an order-driven market. *The Journal of Portfolio Management*, 24(2), 47–55.

Madhavan, A., & Sofianos, G. (1998). An empirical analysis of NYSE specialist trading. *Journal of Financial Economics*, 48(2), 189–210.

Jones, C. M. (2002). A century of stock market liquidity and trading costs. https://ssrn.com/abstract=313681.

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.

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.

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

Mastromatteo, I. (2015). Apparent impact: The hidden cost of one-shot trades. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(6), P06022.

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

High-Frequency Trading and Market-Making

Avellaneda, M., & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3), 217–224.

Jones, C. M. (2013). What do we know about high-frequency trading? https://ssrn.com/abstract=2236201.

Menkveld, A. J. (2013). High frequency trading and the new market-makers. *Journal of Financial Markets*, 16(4), 712–740.

Biais, B., & Foucault, T. (2014). HFT and market quality. *Bankers, Markets & Investors*, 128, 5–19.

Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *Review of Financial Studies*, 27(8), 2267–2306.

Menkveld, A. J. (2016). The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics*, 8, 1–24.

Guéant, O. (2016). Optimal market-making. arXiv:1605.01862.

Jovanovic, B., & Menkveld, A. J. (2016). Middlemen in limit order markets. https://ssrn.com/abstract=1624329.

Menkveld, A. J., & Zoican, M. A. (2016). Need for speed? Exchange latency and liquidity. https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2442690.