On the relationship between accuracy and profitability in over-the-counter market-making

Peter Cotton

November 26, 2024

Abstract

Intuitively, accuracy in microstructure prediction at a granular level must relate to profitability for a market participant, but this is not trivial to formalize. Here we provide an approximation using a stylized steady-state model for over-the-counter trading modeled as a sequence of sealed bid auctions. A simple picture emerges due to a mildly surprising feature of this model: one does not need to know the fair price, only where others are bidding.

1 Main result

The primary purpose of this note is to provide a theoretical model for trading where profitability can be related to accuracy in a simple fashion

$$\texttt{Profitability degradation} = \frac{1}{2}(\texttt{accuracy degradation}) \tag{1}$$

The route to this heuristic, and the definitions of both sides of this equation, is through a relatively unknown model for sequential closed bid auctions where the optimal policy can be determined [1].

We deviate from the market microstructure theory directed at equity markets where central limit order books are the dominant mechanism, because those aren't a good general definition of "trading". Trade in financial markets, not to mention art or baseball cards, usually occurs in a very different way.

One says *usually* in the financial context because the largest markets by volume are in fixed income and currencies. Typically, customers send inquiries to multiple dealers who then compete on price. Messaging networks and request-for-quote protocols represent an electronification of the phone call that is increasingly important, but these largely preserve the same structure.

The essential feature is that the market maker is responding to incoming inquiries and aims to maximize profitability while balancing this against the need to manage inventory. They must compete with other dealers by replying to the client with a good bid or offer. And in this setting, we can exhibit a model where the optimal policy is quite well understood.

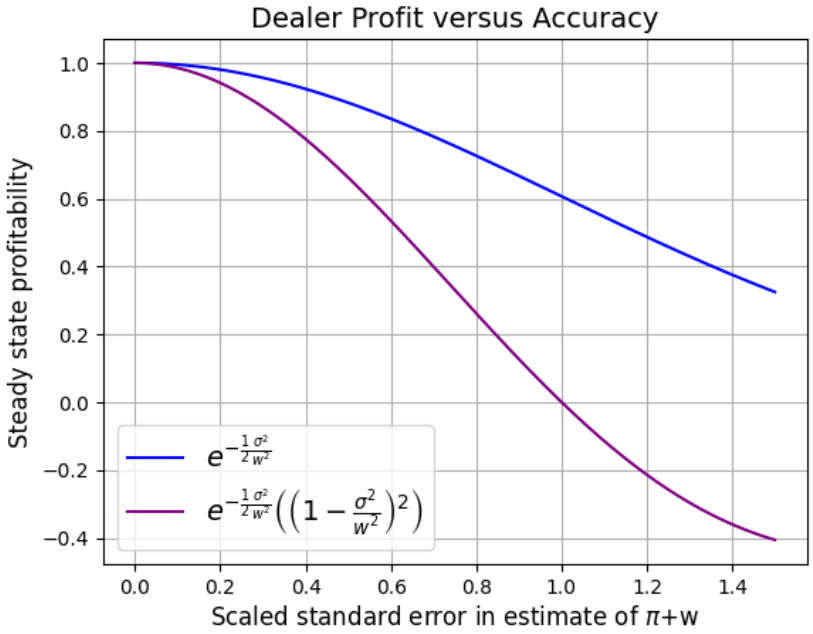

Figure 1: Profitability of a market maker as a function of the accuracy with which they can predict the location parameter for the distribution representing the best competing bid or offer. Here z is the market-width scaled standard error in the estimate of $\hat{y} = \pi + w$ (i.e., the standard deviation in $\hat{y}$ error divided by market width w), and π is the fair market price. The optimal policy does not require knowledge of the location of the fair price itself, which is something of a mathematical accident arising due to the assumption of exponentially distributed markups.

This model, though flawed in its simplicity, allows us to estimate the first-order impact of improved prediction accuracy on market-making profitability, and thus improve our intuition as to what the true relationship might be in reality.

The relationship between accuracy and profit is represented in the heuristic Equation 1 and with more fidelity in Figure 1, where two profitability curves are displayed. One applies to a dealer who either is blissfully unaware of their microstructure prediction inaccuracy, or is choosing to ignore it. The second applies to a dealer who takes their own inevitable errors into account and, in accordance, offers slightly more defensive responses to trade inquiries.

Two brief comments on the units in Figure 1 are in order. First, the x-axis has a straightforward interpretation because the units are in multiples of the market width w, which is a measure of (half) bid-offer spread applicable even in the absence of continuous bids and offers. We shall be defining it with respect

to the distribution of the best bid and offer.

Second, in respect to the y-axis, these approximations are made on a relative scale where unit profitability applies to a market-maker who makes zero forecasting error – though it is important to clarify what variety of omniscience this ideal represents. It is not to imply perfect knowledge of where others will quote, merely an ability to discern without error the location parameter determining the distribution of competitor quotes.

The dealer whose policy we optimize competes with others whose markups and markdowns are exponentially distributed around a commonly discerned mid-price.[1] Perhaps surprisingly, it was shown that $\hat{y} = \pi + w$ is a sufficient statistic and the *only quantity that needs to be predicted*. One does not need to know where the mid is! This simplification allows us to speak of accuracy in the singular, and it leads to two conclusions, one for a dealer who is unaware of their own prediction errors or chooses to ignore them:

$$\text{Relative profitability (bold dealer)} \approx e^{-z^2/2}\left(1 - z^2\right),$$

and another for the more self-aware dealer who will take their own errors into account and back off accordingly, offering slightly wider prices:

$$\text{Relative profitability (humble dealer)} \approx e^{-z^2/2}.$$

where z is an accuracy statistic normalized by the market width, to be defined. In the following sections, we elaborate on the notation and the derivation of these results.

2 Review of the exponential markup model

As explained in more detail in [2], trading is viewed as a sequence of sealed bid auctions. The dealer competes with others whose markups and markdowns are stochastic and exponentially distributed around a commonly discerned mid price.[2] In responding to a request for a quote from a client, the dealer (i.e. market maker) must determine an optimal bid or offer to disseminate by taking several estimates into account.

Those considerations include the obvious linear penalty associated with holding inventory (storage or funding cost); a quadratic penalty that arises when we change reference frame to one where the mid price does not change (a risk term); the less obvious benefit to holding inventory arising from the fact that it increases the optionality of each trading opportunity; the width of the market (inverse hazard rate for the inside market markup distribution); and the errors in predicting the location parameter for competitors' disseminated prices.

[1] This model is not a Nash equilibrium. We play the role of the $j+1$'st dealer who assumes their behavior will not influence that of any other participant.

[2] In this model the optimal policy involves bidding and offering outside the typical market width, as will be elaborated in what follows, so the shape of the inside bid or offer distributions are most relevant at this approximate distance from the mid.

The Bellman equation and optimal steady state behavior of the market maker was discussed in [2] based on previously unpublished research presented in [1]. The key insight is that the location parameter for the inside market distribution is a sufficient statistic, and that to first order there is a simple relationship between the standard error in the estimate of this location and profitability of a market maker.

Although it involves a sign change relative to a Bellman value function, it is most natural for financial participants to cast the Bellman equation in terms of a quantity $\nu(x)$, the maximum markdown a market maker is willing to accept to immediately liquidate an inventory position x Specifically, suppose the market maker has inventory $x > 0$ and the current fair price is p. If a third party offers to buy the entire position at price $p - \nu(x)$, the market maker is indifferent between accepting or declining the offer.

Naturally, it is possible to derive the Bellman equation by standard means but it is also possible to derive conditions $\nu(x)$ in a way that mirrors the control-theoretic logic while maintaining a clear market interpretation. We can consider two paths:

1. The market maker immediately liquidates the position at a cost of $\nu(x)$, and enters the next trading opportunity with zero inventory.

2. The market maker retains the position until the next trading opportunity and then liquidates.

Equating the expected costs of these two strategies leads to a consistency condition that relates $\nu(x)$ to the expected profits from future trading opportunities, as shown in [2]. From this financial argument, we derive the break-even markups and markdowns when the market maker is indifferent between trading or not trading. For buying opportunities when a client wishes to sell a size s, we can define the break-even markdown:

$$K^{\uparrow}(x; s) = \epsilon + \frac{\nu(x + s) - \nu(x)}{s}, \tag{2}$$

which is the amount the dealer would subtract from the fair market price so as to be completely indifferent to whether their bid is lifted or not. Similarly for selling opportunities:

$$K^{\downarrow}(x; s) = \epsilon + \frac{\nu(x - s) - \nu(x)}{s}, \tag{3}$$

Here the newly introduced quantity ϵ represents adverse selection costs. The optimal choice can be shown to be:

$$m^{\text{ask}}(x; s) = \frac{1}{h(m^{\text{ask}}(x; s))} + K^{\downarrow}(x; s), \tag{4}$$

and similarly for buying:

$$m^{\text{bid}}(x; s) = \frac{1}{h(m^{\text{bid}}(x; s))} + K^{\uparrow}(x; s), \tag{5}$$

where $h(\cdot)$ is the hazard rate of the competitors' markup distribution. These results are true for any assumption about the distribution of the best bids and offers, but they simplify if the hazard rate h is constant and the inside market is exponentially distributed. In that special case where the optimal markups simplify to:

$$m^{\text{ask}}(x;s) = \max\left(\frac{1}{h} + \epsilon + \frac{\nu(x-s)-\nu(x)}{s},\ 0\right), \tag{6}$$

$$m^{\text{bid}}(x;s) = \max\left(\frac{1}{h} + \epsilon + \frac{\nu(x+s)-\nu(x)}{s},\ 0\right). \tag{7}$$

These expressions decompose the optimal markup into three components:

1. A base markup determined by what we term the market width $w = \frac{1}{h}$.
2. An adverse selection component, ϵ.
3. A marginal inventory cost component, $\frac{\nu(x\pm s)-\nu(x)}{s}$.

The *true* inventory cost ν is related to, but not equivalent to, the obvious inventory cost $c(x)$ that is modeled by a term linear in $|x|$ and a risk term proportional to x^2.

3 Profitability when errors are not ignored

Though exponential markups are a stylized assumption, they convey an important intuition about the importance of knowing the other dealers' markups versus the importance of knowing where the mid is.

Let π denote the fair market (mid) price, h the hazard rate for the inside market and $w = 1/h$ the corresponding market width. Assume the market maker's estimates for both mid price and market width are normally distributed:

$$\begin{aligned} \hat{\pi} &= \pi + \epsilon_\pi \\ \hat{w} &= w + \epsilon_w \end{aligned}$$

where $\epsilon_\pi \sim N(0, \sigma_h^2)$ and $\epsilon_w \sim N(0, \sigma_w^2)$ are i.i.d. - an assumption we will return to. In what follows we could use $\eta \in \{\uparrow, \downarrow\}$ to keep the discussion general but for concreteness, consider the case $\eta = \downarrow$ when the market maker has an opportunity to sell.

Claim 1. *The location of the best response by competitors, equal to $\hat{\pi} + \hat{w}$, is a sufficient statistic summarizing π and w.*

Proof: We have already seen that if the market maker knows π and h exactly then her choice for marked up price $\pi^\downarrow = \pi + m^\downarrow$ will maximize the gain

$$G(\pi^\downarrow) = F(\pi^\downarrow)\left(m^\downarrow - K^\downarrow(x;s)\right)$$

where $F(\pi^{\downarrow}) = e^{-h(\pi^{\downarrow}-\pi)}$. Assuming we are in the region where the first order condition is relevant, the best choice $\pi^{ask} = \pi + m^{ask}$ will satisfy

$$\overbrace{\pi^{ask} - \pi}^{m^{\downarrow}} - K^{\downarrow}(x;s) = w$$

with gain

$$\begin{aligned} G(\pi^{ask}) &= e^{-h(\pi^{ask}-\pi)} \left(\pi^{ask} - \pi - K^{\downarrow}(x;s)\right) \\ &= we^{-h(\pi^{ask}-\pi)} \ . \end{aligned}$$

On the other hand if the market maker knows h and π only approximately, with estimates $\hat{h}$ and $\hat{\pi}$ respectively, she will make a different choice π_{ϵ}^{ask} solving the wrong problem. Her choice instead will satisfy

$$\pi_{\epsilon}^{ask} - \hat{\pi} - K^{\downarrow}(x;s) = \hat{w} \ ,$$

and we observe, by subtraction, that this diverges from the optimal choice:

$$\begin{aligned} \pi_{\epsilon}^{ask} - \pi^{ask} &= \hat{w} + \hat{\pi} + K^{\downarrow}(x;s) - \pi - w - K^{\downarrow}(x;s) \\ &= \overbrace{\hat{w} - w}^{\epsilon_w} + \overbrace{\hat{\pi} - \pi}^{\epsilon_\pi} \\ &:= \epsilon_{w+\pi} \\ &\sim N(0, \sigma^2) \ , \end{aligned}$$

showing, incidentally, that the errors in both price and width translate directly into error in the optimal choice of response to the request for a quote (RFQ). The sum of errors is also the error in the market maker's location estimate for the mean of the best price response by his competitors. Thus $\hat{\pi}+\hat{w}$ is a sufficient statistic.

From the proof of Claim 1 we let $G(\pi)$ denote the expected net benefit to the market maker of the trading opportunity; now we let $E^{\sigma^2}[G(\pi_{\epsilon})]$ the same under uncertain knowledge of w and h and define **the relative efficiency:**

$$R := \frac{E^{\epsilon}[G(\pi_{\epsilon}^{ask})]}{G(\pi^{ask})} \ .$$

Claim 2. *Let $\sigma^2 := \sigma_{\pi}^2 + \sigma_w^2$, i.e. the combined variance of $\hat{\pi} + \hat{w}$. If $h = 1/w$ is small (relative to σ^2), then the relative efficiency is approximated by*

$$R \approx \bar{R}_0 = e^{-\frac{1}{2}\frac{\sigma^2}{w^2}} \left(1 - \frac{\sigma^2}{w^2}\right) \ . \tag{8}$$

Proof: If $\pi_\epsilon > \pi$, then the gain uncertainty markup is

$$\begin{aligned} G(\pi_\epsilon^{ask}) &= e^{-h(\pi_\epsilon^{ask}-\pi)}\left(\pi_\epsilon^{ask} - \pi - K^{\downarrow}(x;s)\right) \\ &= e^{-h(\pi^{ask}-\pi+\epsilon_{w+\pi})}\left(\overbrace{\pi^{ask} - \pi - K^{\downarrow}(x;s)}^{=w} + \epsilon_{w+\pi}\right) \\ &= we^{-h(\pi^{ask}-\pi)}e^{-h\epsilon_{w+\pi}}\left(1 + \frac{\epsilon_{w+\pi}}{w}\right) \\ &= G(\pi^{ask})e^{-h\epsilon_{w+\pi}}\left(1 + \frac{\epsilon_{w+\pi}}{w}\right) \ . \end{aligned} \tag{9}$$

A somewhat more careful calculation takes into account the possibility that due to errors in estimation the market maker might offer through the mid π and, in that event, always do the trade. In otherwords, if $\pi_\epsilon \leq \pi$ then the dealer automatically makes the trade, and the gain is

$$G(\pi_\epsilon^{ask}) = \pi_\epsilon^{ask} - \pi - K^{\downarrow}(x;s) = w + \epsilon_{w+\pi} \ . \tag{10}$$

Combining equations (9) and (10), we have the efficiency ratio:

$$\begin{aligned} R &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-(\pi^{ask}-\pi)}^{\infty} e^{-\frac{1}{2\sigma^2}\epsilon^2}e^{-h\epsilon}(1+h\epsilon)d\epsilon \\ &\quad + \frac{1}{G(\pi^{ask})\sqrt{2\pi\sigma^2}}\int_{-\infty}^{-(\pi^{ask}-\pi)} e^{-\frac{1}{2\sigma^2}\epsilon^2}\left(w+\epsilon\right)d\epsilon \ . \end{aligned}$$

If h is small then w is large, and assuming σ^2 is not too large we have

$$\pi^{ask} - \pi \gg 1 \ ,$$

and then the chance of offering below the mid is small, which prompts us to define an approximate efficiency ratio

$$\bar{R}(\epsilon_\pi, \epsilon_w) := e^{-\frac{\epsilon_{w+\pi}}{w}}\left(1 + \frac{\epsilon_{w+\pi}}{w}\right) \ , \tag{11}$$

where again, $\epsilon_{w+\pi}$ is the difference between the estimated location of the mean of the competitors' best ask and the ground truth. Integrating directly we have, approximately:

$$\begin{aligned} \bar{R}_0 &= E^\epsilon R(\epsilon_\pi, \epsilon_w) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(\epsilon)^2}e^{-h\epsilon}(1+h\epsilon)d\epsilon \\ &= e^{\frac{1}{2}\frac{\sigma^2}{w^2}}\left(1 - \frac{\sigma^2}{w^2}\right) \ . \end{aligned}$$

as claimed.

4 Profitability when errors are accounted for

As (11) warns, assuming the correctness of point estimates might not be the best way for the market maker to choose m^{ask}. Instead, she should acknowledges her errors in $\hat{\pi}$ and $\hat{w}$ and act defensively.

Claim 3. *Expected efficiency is maximized by replacing the point estimate $\hat{\pi}+\hat{w}$ with a cautious estimate $\hat{\pi}+\hat{w}+h\sigma^2$ whereupon the mean efficiency is approximated by*

$$\bar{R}_{h\sigma^2} = e^{-\frac{1}{2}\frac{\sigma^2}{w^2}} \ . \tag{12}$$

Proof: If the market maker biases her estimate of $\hat{\pi}+\hat{w}$ by an amount φ then $\epsilon_{\pi+w}$ will be normally distributed around φ with variance equal to her error σ^2. As with (9) her mean efficiency must be

$$\begin{aligned} \bar{R}_{\varphi} &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(\epsilon-\varphi)^2} e^{-h\epsilon}(1+h\epsilon)d\epsilon \\ &= e^{\frac{h^2\sigma^2}{2}} e^{-h\varphi}(-h^2\sigma^2+h\varphi+1) \end{aligned} \tag{13}$$

Let $\upsilon(\varphi) := e^{-h\varphi}(-h^2\sigma^2+h\varphi+1)$. Since $\upsilon'(\varphi) = \left(-h+\frac{h}{-h^2\sigma^2+h\varphi+1}\right)\upsilon(\varphi)$ we set the term in the denominator equal to 1 to maximize efficiency whereupon $\varphi = h\sigma^2$ and

$$\bar{R}_{\varphi} = e^{-\frac{h^2\sigma^2}{2}} \ ,$$

as claimed.

There are two flaws with this approximation. First, we can only apply an offset of $\hat{h}\sigma^2$ not $h\sigma^2$ since the latter is unknown. Second, our approximation assumes no chance of offering through the true mid. However if $\pi_\epsilon < \pi$ the gain (in fact loss) is actually

$$G(\pi_\epsilon^{ask}) = \pi_\epsilon^{ask} - \pi - K^{\downarrow}(x;s)$$

because we always trade so (9) is not valid for all π_ϵ^{ask} and either is (11) for all $\epsilon_{\pi+w}$. In practice then, a market maker might want to deviate from (13) especially when given an opportunity to get out of a large inventory. The optimal efficiency would maximize the exact formula

$$R_{\varphi} = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(\epsilon-\varphi)^2} \min\left(e^{-h\epsilon}, 1\right)(1+h\epsilon)d\epsilon \ ,$$

We emphasize that Claim 3 is merely a heuristic and not an exact solution.

References

[1] Peter Cotton and Andrew Papanicolaou. Trading illiquid goods: Analytic results and intuition, May 2017. Presented at Intech Investments and NYU Tandon.

[2] Peter Cotton and Andrew Papanicolaou. Trading illiquid goods: Market making as a sequence of sealed-bid auctions, with analytic results, April 2022. Accessed from GitHub repository.