Mathematics in Economics: Market Microstructure Theory

MAS491 Final Presentation

Ko Sunghun

KAIST

December 10, 2024

Table of Contents

In this homework presentation, we have done a case study of application of mathematics in market microstructure thoery, which leverages game thoery and probability theory to provide insights on complex economic phenomena.

1. Motivation
2. Continuous Market Model without Latency
3. Continuous Market Model with Latency
4. Frequent Batch Auctions as alternative
5. Conclusion

Motivation

High-frequency Trading Arms Race

It is well known that HFT firms compete against each other for faster execution. In the study of Budish, Cramton, and Shim (2015), they studied equilibrium on exchange between these HFT firms under various order priority rules, suggesting alternative market design that potentially reduces spread while effectively terminating endless arms race for speed.

Continuous Market Model without Latency

Evolution of Price

Our goal is to show how endless competition for latency arbitrage is "built-in" in continuous first-come-first-serve (FCFS) orderbook model. To do so, assume following very simple model:

- There is a security X that is traded in continuous market.
- The price of X is perfectly correlated to publicly observable signal y.
- E.g. X being SPY (S&P 500 ETF), y being S&P 500 index.
- y follows stochastic process, which is, compound Poisson jump process with arrival rate λ and jump distribution F. Here we additionally assume F is symmetric and zero-mean.
- Define $J := |F|$, the jump size distribution.

Continuous Market Model without Latency

Players

There are following players in the market: HFT firms and Investors. We first introduce the HFT firms' characteristics.

- They have no intrinsic demand on security X. They just want to provide liquidity to market and earn from spread or snipe the stale quote. To do so, they can either provide liquidity by submitting limit orders, expecting investors to cross the spread, or take stale quotes from other HFT firms.
- For simplicity, they are risk neutral. Their goal is to buy X whenever it is underpriced (i.e. $y > p_t$) and sell X whenever it is overpriced (i.e. $y < p_t$). Here, p_t is price at trading time t. Then objective becomes to maximize the expected profit per unit time.
- We first start with the case where the number of HFT firms, N, is given exogenously. Later, after introducing latency and speed cost c_{speed}, we will consider the case where N is endogenously determined.

Continuous Market Model without Latency

Players

Next we introduce the Investors.

- Investors represents the rest of the market who are insensitive to small fluctuations in execution price and time. They are assumed to arrive market and submit a unit amount of market orders, with arrival rate $\lambda_{\text{investor}}$. By half probability they submit buy order, and vice versa.

Continuous Market Model without Latency

Players

We fix player 1 to be monopolistic liquidity provider (LP). We begin with simpler case where there is no latency. That is, whenever quote from HFT firm is stale, other firms will try to snipe it while the owner of stale quote will try to cancel it. Our goal is to find nash equilibrium in this setting and how spread s is determined. For this setting to be equilibrium, the expected profit from being LP and sniper should be same. Each player's strategy is as follows:

- LP: Maintain a bid and ask spread centered around y, with spread s, expecting investors to cross the spread. If the quote becomes stale from jump of y, try to cancel it.
- Sniper: Submit market order whenever quote becomes stale.

Continuous Market Model without Latency

Equilibrium in No-latency setting

Return from being LP is:

$$\lambda_{\text{investor}} \cdot \frac{s}{2} - \lambda \cdot \mathbb{P}(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N} \quad (1)$$

Meanwhile, return from being Sniper is:

$$\lambda \cdot \mathbb{P}(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N} \quad (2)$$

Here, $\frac{N-1}{N}$ comes from the fact that there are $N-1$ other HFT firms that can snipe the stale quote (orders that arrived at same time are executed in random order). Equating these two, we can find the spread s that maximizes the expected profit of LP. That is, s should satisfy:

$$\lambda_{\text{investor}} \cdot \frac{s}{2} = \lambda \cdot \mathbb{P}(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \quad (3)$$

Continuous Market Model without Latency

Interpretation of Equilibrium

Equation (3) contains implies important insight – the spread times arrival rate of investors, which is, cost of investors per time, is equal to the sum of expected return of sniping and the cost of being liquidity provider. This implies that competition among HFT firms manifests as costs borne by investors. In other words, HFT firms are effectively passing on the costs of competition to the rest of the market participants, including retail investors. In the next model, which incorporates latency, this becomes even more apparent.

Continuous Market Model With latency

Players

While the setting on investors remains same, this time we introduce latency in observing y and cost c_{speed} to reduce such latency. Without paying additional cost c_{speed}, HFT firms can only observe y with delay δ_{slow}. With c_{speed}, they can reduce this delay to δ_{fast}. Define $\delta := \delta_{\text{slow}} - \delta_{\text{fast}}$.

Continuous Market Model With latency

Equilibrium in the presence of Latency

It is clear that HFT firm who didn't pay c_{speed} for additional speed (i.e., lower latency) can never win competition and will be obsolete, resulting in not participating game. In this setting, the expected profit of LP is:

$$\lambda_{\text{investor}} \cdot \frac{s}{2} - \lambda \cdot \mathbb{P}(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N} - c_{\text{speed}} \geq 0 \tag{4}$$

Meanwhile, the expected profit of Sniper is:

$$\lambda \cdot \mathbb{P}(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N} - c_{\text{speed}} \geq 0 \tag{5}$$

where s is spread and N is the number of HFT firms who paid c_{speed} to participate in the game.

Continuous Market Model With latency

Interpretation

Equating these two with profitability condition, one may find s^* and N^* that satisfy:

$$\lambda_{\text{investor}} \cdot \frac{s^*}{2} = \lambda \cdot \mathbb{P}(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2}) \tag{6}$$

$$\lambda_{\text{investor}} \cdot \frac{s^*}{2} = N^* \cdot c_{\text{speed}} \tag{7}$$

Again, from (7), it becomes clear that the cost of competition in HFT firms is effectively passed on to investors in the form of spread.

Frequent Batch Auctions as alternative

Frequent Batch Auctions

The authors of referenced paper suggest alternative market design called frequent batch auctions. In this market design, orders are batched for a short period of time, and then executed in uniform price. Throughout the following slides, we will see how this market design can potentially reduce spread and effectively eliminate the need for endless competition for speed. Informally, their idea is that by setting batch interval long enough, they enables liquidity providers to cancel their stale quotes whenever huge enough jump occurs, unless it happens at the very end of batch. This will reduce the cost of being LP and in conclusion, reduce the spread.

Frequent Batch Auctions as alternative

Equilibrium in Batch Auctions

Note that in this setting, limit orders are not filled in FCFS but rather in pro-rata manner. That is, each limit order at same price level will be filled in proportion to their size. In setting such that batch interval τ is long enough than δ, benefit from lower latency is negligible and no firm pay for additional speed. Moreover, since there is effectively no risk of being sniped, now HFT firms compete against each other to "provide" liquidity at zero bid-ask spread, not sniping it. Now let Q^* be the size of limit order from LP(s). For this condition to be hold, as forementioned, paying c_{speed} should not be profitable. That is, profit from additional speed should be less than or equal to c_{speed}. Formalizing this, we have:

$$\lambda \cdot \frac{\delta}{\tau} \cdot \mathbb{E}(J) \cdot Q^* \leq c_{\text{speed}}. \tag{8}$$

Frequent Batch Auctions as alternative

Interpretation

Thus for any given setting on price movement, latency gap and cost for reducing latency, by setting tau long enough, one can create market with zero spread and depth of Q^*. This is in stark contrast to the continuous market model, where spread is determined by the cost of competition among HFT firms and always strictly positive. Authors concluded with suggesting, from real-world statistics, setting interval around 100 milliseconds to 1 second will be enough to achieve this.

Conclusion

Conclusion

In this presentation, we have seen how market microstructure theory, which leverages game theory and probability theory, can provide insights on complex economic phenomena. In particular, we reviewed the paper of Budish, Cramton, and Shim on how it is important to design market structure correctly, by showing contrasting results between continuous and batch execution of orders. Like this example, mathematics can be a powerful tool in understanding and designing economic systems.