High Frequency Market Making

Yacine Aït-Sahalia  Mehmet Saglam

Princeton University and NBER  Princeton University

The Evolving Structure of the U.S. Treasury Market
Federal Reserve Bank of New York
October 20-21, 2015
1. A Model of HFMM

• Premise: Compared to traditional market makers

  – HFMMs are better informed than their counterparties: able to extract signals about the direction of the order flow

  – And are faster

• What can we expect when HFMMs become the primary providers of liquidity?
• Inventory discipline is the primary means of risk control by the HFMM who is risk-neutral, but penalized for holding inventory.

• LFTs are randomly arriving noise traders submitting market orders.

• The HFMM posts quotes and aims to capture the spread as often as possible.

• The HFMM receives a signal that is informative, but not perfect, about the sign of the incoming market order from LFTs.

• Optimal Policy: HFMM always quote unless inventory thresholds are exceeded.

• When deciding whether to quote or not, the HFMM is constantly weighing the potential of capturing the spread vs. the cost of increasing his inventory.
2. Predictions of the Model

- Objective value and optimal inventory limits as a function of model parameters
  - the arrival rate of the LFTs, $\lambda$
  - the arrival rate of the HFMM’s signal, $\mu$
  - the accuracy of the signal, $p$
  - the bid-offer spread, $c$
  - the coefficient of inventory aversion, $\gamma$
2.1. LFTs’ Market Orders Arrival Rate

Optimal value and inventory trading limits

![Graph showing the relationship between arrival rate of market orders and objective value/critical limits.](image-url)
2.2. HFMM’s Signal Arrival Rate (or Latency)

Optimal value and inventory trading limits

![Graph showing optimal value and inventory trading limits]
2.3. Signal Accuracy

Optimal value and inventory trading limits
2.4. Bis-Ask Spread

Optimal value and inventory trading limits

![Graph of Objective Value and Critical Limits vs. Bid-Offer Spread](image)
2.5. Inventory Aversion

Optimal value and inventory trading limits

![Graph showing Inventory Aversion (\(\Gamma\)) vs. Objective Value and Critical Limits vs. Inventory Aversion (\(\Gamma\)).]
2.6. Provision of Liquidity by the HFMM

Long-run Probability of LFTs’ Orders Being Filled by the HFMM
2.7. Endogenous Cancellations by the HFMM

Long-run probability of an existing quote being canceled by the HFMM
3. Price Volatility

- Add price variability in the form of jumps in the asset’s fundamental value.

- The HFMM has no informational advantage regarding these price movements; his only signal is about the likely direction of the order flow.

- Volatility introduces adverse selection: the HFMM may get stuck with stale quotes that can be sniped by another HFT.
Example: A Simulated Path with Volatility
Long-run probability of quoting as a function of the price jump arrival rates
• When the price is more volatile, the likelihood that the HFMM will provide liquidity decreases.

• This is because this volatility introduces a new source of risk for the HFMM (excess inventory) that is not compensated for and for which he holds no advantage (no signal).

• So while the HFMM provides plenty liquidity in normal times, it is optimal for the HFMM to withdraw when the market needs that liquidity the most...
4. Competition Among HFMMs

- Duopoly: Splitting the Rent

Optimal value achieved by the HFMM: Monopoly vs. Duopoly
• The rent extracted from LFTs gets split between the two market makers.

• The faster the HFMM, the more of the rent he is able to capture: there are benefits to becoming faster among HFMMs.

• LFTs are better off when market makers compete compared to the monopolistic HFMM situation.
5. Comparing Different HFT Regulations

- Three policies in the context of the model: imposing a transaction tax on each trade, setting minimum rest times on limit orders and taxing cancellations of limit orders.

- Objective: induce the HFMM to provide liquidity that is more resilient to increases in volatility = procyclical with respect to volatility
  - We find that none of the three policies result in an improvement compared to doing nothing.
  - Transaction taxes result in less liquidity both in low and high volatility environments.
  - Both minimum rest times and a cancellation tax result in more liquidity in good (low volatility) environments but less in bad (high volatility) environments = countercyclical.
5.1. Tobin Tax: Taxing Transactions

- Equivalent to a reduction in the spread. Transaction taxes reduce the incentive to quote.
5.2. Minimum Rest Time

- Mandatory rest times increase the provision of liquidity when volatility is low, but decrease it when volatility is high.
5.3. Taxing Order Cancellations

- Tax the HFMM whenever an existing quote is cancelled.

- Cancellation taxes encourage the HFT to quote more when volatility is low but less when it is high.
6. Conclusions

- The latency advantage of a HFT can be quantified in a fully optimizing model.

- Predictions of the model:
  - The HFMM trades often, carries little inventory, captures the spread from LFTs.
  - Lower latency is beneficial to the HFMM.
  - Order cancellations occur endogenously in the model.
  - In good times, the HFMM improves liquidity. But when price volatility increases, the HFMM decreases his liquidity provision.
  - Competition among HFMMs lead to splitting the rent and benefits LFTs.
• Regulations?

  – Taxing transactions is ineffective: it uniformly reduces the provision of liquidity

  – Mandatory rest times and cancellation taxes increase the provision of liquidity when volatility is low

  – But decrease it when volatility is high

  – So both fail to encourage countercyclical liquidity provision.
Details?