Frontiers in Emerging Multidisciplinary Sciences (Open Access)

Volume 02, Issue 12, December 2025,
Publish Date: 31-12-2025
PageNo.24-28

Dynamic Liquidity Provision and Risk Internalization in Fragmented High-Frequency Markets: A Comprehensive Framework for Algorithmic Market Making and Execution

Drek Grey
Department of Financial Mathematics, London School of Economics and Political Science, United Kingdom

ABSTRACT

This research article provides an extensive investigation into the evolution of algorithmic market making and liquidity provision within the contemporary financial landscape, characterized by high-frequency trading (HFT) and over-the-counter (OTC) decentralization. By synthesizing foundational theories of information asymmetry with modern stochastic control models, this paper examines how market participants navigate the dual challenges of inventory risk and adverse selection. The study explores the shift from traditional bid-ask spread models to complex internalization strategies used by electronic dealers, particularly in Foreign Exchange (FX) and cryptocurrency-native environments. Through a detailed descriptive analysis of market microstructures, we evaluate the impact of latency, "last look" execution protocols, and the "arms race" for speed on overall market quality. The findings suggest that while high-frequency intermediaries provide essential liquidity, the concentration of flow within internalized pools creates a bifurcated market structure that necessitates sophisticated dimensionality reduction and closed-form approximation techniques for risk management. The article further discusses the implications of stochastic liquidity demand and the role of signal-based learning in optimizing execution, providing a robust theoretical framework for future regulatory and institutional developments.

KEYWORDS: Market Making, High-Frequency Trading, Liquidity Internalization, Inventory Risk, Information Asymmetry, Algorithmic Execution, Stochastic Control.

INTRODUCTION

The architectural transformation of global financial markets over the last four decades has been defined by the transition from human-intermediated floor trading to a hyper-automated, fragmented ecosystem driven by sophisticated algorithms. At the heart of this evolution lies the role of the market maker-the entity tasked with providing continuous liquidity by quoting both buy and sell prices. Historically, the compensation for this service was the bid-ask spread, a simple compensation for the risks of holding inventory and the potential for being "picked off" by better-informed traders (Copeland & Galai, 1983). However, in the modern era, the simplicity of the bid-ask spread has been replaced by a multifaceted optimization problem involving microsecond latency, stochastic price jumps, and the strategic internalization of order flow.

The problem statement of this research centers on the increasing complexity of liquidity provision in environments where information is asymmetric and execution is near-instantaneous. As high-frequency traders (HFTs) take advantage of speed to exploit price discrepancies across venues (Aït-Sahalia et al., 2013), traditional market makers must adapt their strategies to avoid significant losses. This is particularly evident in the Foreign Exchange (FX) and Over-the-Counter (OTC) markets, where the lack of a centralized exchange allows for varied execution protocols, including the controversial "last look" mechanism (Cartea, Jaimungal, & Walton, 2019). The literature gap addressed here is the need for a unified theoretical framework that connects the micro-level behavior of individual algorithms with the macro-level stability and quality of the market.

Early theories of market making focused primarily on the informational component of the spread. It was argued that the spread exists because market makers face two types of traders: liquidity traders, who trade for exogenous reasons, and informed traders, who possess superior knowledge about the future price of an asset (Copeland & Galai, 1983). In this context, the market maker is essentially granting a free option to the informed trader. If the market maker’s quote remains static while the informed trader knows the price is about to shift, the market maker suffers an adverse selection cost. Modern research has expanded this to include high-frequency data, demonstrating that the risk-return trade-off for these intermediaries is highly sensitive to the frequency and granularity of the data used for decision-making (Bali et al., 2006).

Furthermore, the rise of "cluster trading" and the presence of high-frequency "arms races" have introduced new dynamics into market efficiency. Budish et al. (2015) highlighted that the continuous-time nature of modern exchanges may be fundamentally flawed, as it encourages a race for speed that does not necessarily improve market depth but rather redistributes wealth among high-speed participants. This leads to a situation where proprietary and buy-side algorithmic traders engage in a metaphorical game of "hide-and-seek," where the former attempts to predict and front-run the large, slow-moving orders of the latter (Arumugam et al., 2021). The resulting causal links with market quality are complex, often resulting in periods of high liquidity punctuated by sudden "flash" illiquidity.

This article proceeds by detailing the methodologies employed in modern market-making algorithms, focusing on stochastic control and dimensionality reduction. We then analyze the results of these strategies in the context of FX and cryptocurrency markets, discussing the critical role of internalization and the theoretical implications of these findings for the future of financial stability.

METHODOLOGY

The methodology underpinning modern algorithmic market making is rooted in the mathematical discipline of stochastic control, specifically the optimization of a value function that represents the expected utility of a market maker's terminal wealth. Unlike traditional investment strategies that seek to maximize returns from price appreciation, a market maker seeks to maximize the accumulated spread while minimizing the variance of their wealth caused by inventory fluctuations. This section elaborates on the text-based conceptualization of these mathematical frameworks without the use of formal equations, focusing on the logic and theoretical structures defined in the works of Cartea, Jaimungal, and others.

The core of the market-making model involves a continuous-time Markov process where the mid-price of an asset follows a stochastic path, often modeled as a Brownian motion or a jump-diffusion process. The market maker must decide, at every instant, where to place their bid and ask quotes relative to this mid-price. These quotes are not static; they are dynamic offsets that respond to two primary inputs: the current inventory level and the estimated intensity of arriving buy and sell orders. If a market maker has a long inventory (meaning they have bought more than they have sold), they are exposed to the risk of a price drop. To mitigate this, they will theoretically lower both their bid and ask quotes to attract more sellers and discourage buyers, thereby "leaning" their quotes to return to a neutral inventory position (Cartea, Jaimungal, & Penalva, 2015).

A significant advancement in this methodology is the incorporation of model uncertainty. In real-world markets, the parameters of the price process and the arrival rate of orders are not known with certainty. Market makers must therefore build algorithms that are robust to "ambiguity" or model misspecification. Cartea, Donnelly, and Jaimungal (2017) suggest that an optimal strategy must account for the worst-case scenario within a set of plausible models, ensuring that the market maker remains profitable even when their primary assumptions about market volatility or liquidity demand are slightly inaccurate. This is often achieved through a penalty term in the optimization function that discourages strategies that are overly sensitive to parameter changes.

In the OTC and FX markets, the methodology shifts toward internalization. Internalization is the process where a dealer matches a client's buy order with another client's sell order internally, rather than hedging the net position on an external, public exchange. This allows the dealer to capture the full bid-ask spread without paying the transaction costs associated with public venues (Butz & Oomen, 2019). To model this, researchers use "skewing" techniques, where the dealer's quotes are adjusted based on the probability of receiving an offsetting order within a certain time window. The complexity here lies in the "multi-asset" nature of the problem. A dealer in FX does not just manage one currency pair but dozens. Bergault et al. (2021) demonstrate that when managing multiple assets, the "curse of dimensionality" makes it impossible to solve the optimization problem using traditional methods. Instead, they propose dimensionality reduction techniques and closed-form approximations. These approximations allow the algorithm to compute "near-optimal" quotes in real-time by simplifying the relationship between correlated assets, such as the Euro and the British Pound.

Another critical methodological component is the use of signals and learning. Modern market makers do not just look at their own inventory; they monitor the entire market "limit order book" (LOB) for signals of impending price changes. For instance, an imbalance in the volume of orders at the best bid versus the best ask can be a strong predictor of a short-term price move. Drissi (2022) explores how market makers can integrate these signals into their stochastic control framework. By using machine learning or quantized forward distributions, the algorithm can "learn" the distribution of future price moves and adjust its quotes preemptively (Ceffer et al., 2018). This transforms market making from a reactive inventory-management task into a proactive predictive-trading task.

Finally, in the context of decentralized and crypto-native markets, the methodology must account for unique risks such as "blockchain latency" and "smart contract execution risk." Kale (2025) discusses FX hedging algorithms designed specifically for crypto-native companies, where the volatility is significantly higher than in traditional fiat markets. The methodology involves using "stablecoins" as intermediary assets to hedge exposure rapidly across different liquidity pools, requiring a high degree of automation and integration with decentralized finance (DeFi) protocols.

RESULTS

The descriptive analysis of findings from the implementation of these high-frequency and market-making strategies reveals a nuanced picture of modern liquidity. One of the most prominent results is the effectiveness of internalization in reducing the "market impact" of trades. When electronic FX dealers internalize flow, they prevent the information contained in that flow from reaching the public market immediately. As Butz and Oomen (2019) observe, this leads to a reduction in the "decay" of the spread, meaning that the prices quoted to clients remain more stable. However, the results also show that this creates a "winner-takes-all" dynamic. Large dealers with massive client bases can internalize more flow, allowing them to offer tighter spreads than smaller competitors who must hedge on public exchanges and incur higher costs (Bergault & Guéant, 2021).

In terms of the "size matters" hypothesis, the data indicates that in OTC markets, the ability to handle large trade sizes is a significant competitive advantage. Market makers who can accurately price large blocks of assets using dimensionality reduction techniques are able to capture a larger share of the institutional market. The results of using closed-form approximations (Bergault et al., 2021) show that these mathematical shortcuts are highly effective; they allow for quote updates that are nearly as accurate as those derived from full-scale simulations but are computed in a fraction of the time. This speed is essential because, as Aït-Sahalia et al. (2013) demonstrate, even a few microseconds of delay can result in a market maker being "sniped" by a faster participant.

Regarding high-frequency data in the options market, Andersen et al. (2021) find that the descriptive characteristics of trades and quotes are highly seasonal and sensitive to macroeconomic announcements. The results show that during periods of high volatility, market makers significantly widen their spreads or withdraw from the market entirely. This "liquidity mirage" is a recurring finding in high-frequency research: the liquidity that appears to be present during calm periods often vanishes the moment it is most needed. This is corroborated by Benos et al. (2017), who studied the interactions among high-frequency traders and found that their strategies are often highly correlated. When one HFT algorithm detects a risk and pulls its quotes, others often follow suit instantaneously, leading to a "liquidity crater."

The study of information asymmetry in the Chinese stock market by Hu et al. (2019) provides further evidence of the results of cluster trading. Their findings indicate that when informed traders move in clusters, market efficiency actually decreases in the short term because the market maker’s "learning" process is overwhelmed by the sheer volume of biased flow. Similarly, the behavior of institutional investors in the Taiwan stock index futures market shows that their trading behavior is a leading indicator of future returns, but only when the market maker is unable to distinguish between informed and uninformed flow (Lai et al., 2014).

In the FX market specifically, the result of the "last look" protocol is particularly contentious. Cartea, Jaimungal, and Walton (2019) find that the ability of a liquidity provider to reject a trade after the client has accepted the quote provides the provider with a significant "risk-free" advantage. This results in lower costs for the provider but can lead to higher "rejection rates" for clients, particularly during volatile periods. This suggests that the "visible" spread in FX might be misleading, as the "effective" spread must account for the probability of trade rejection.

Finally, the results of the total cost of transactions on major exchanges like the NYSE indicate that despite the decrease in explicit commissions over the decades, the "hidden" costs-such as market impact and the cost of being picked off-remain substantial (Berkowitz et al., 1988). Modern algorithmic traders have successfully shifted the cost structure, but the total economic friction of trading remains a persistent challenge for large institutional buy-side firms.

DISCUSSION

The deep interpretation of these results suggests that we are witnessing a fundamental shift in the definition of "market quality." Traditionally, liquidity was measured by the depth of the order book and the narrowness of the spread. However, the findings discussed above indicate that in a world of high-frequency internalized flow, these metrics may be insufficient. The "internalization" of flow means that the most "toxic" (informed) flow is often what ends up on public exchanges, while "passive" (retail) flow is captured privately. This creates a fragmentation that could potentially destabilize the price discovery process. If the public price is only based on a subset of the most aggressive and informed trades, it may become more volatile and less representative of the true consensus value.

The theoretical implications of the "arms race" for speed (Budish et al., 2015) suggest a misalignment of incentives. While speed allows for faster price correction, the marginal benefit to society of moving from millisecond to microsecond execution is questionable, especially when it comes at the cost of massive investment in specialized infrastructure that could be used for other productive purposes. This "positional arms race" creates a barrier to entry, concentrating market-making power into the hands of a few technologically elite firms. This concentration of power leads back to the "size matters" findings of Bergault and Guéant (2021); the scale required to be a modern market maker is now so vast that it may limit the resilience of the market by reducing the diversity of participants.

A significant limitation of current market-making models, including those discussed by Cartea et al. (2015), is their reliance on the assumption of "stationarity" in market conditions. Most models assume that the parameters governing price moves and order arrivals change slowly enough to be estimated. However, in "black swan" events or flash crashes, these parameters can shift instantly and violently. The "model uncertainty" framework (Cartea, Donnelly, & Jaimungal, 2017) is a step toward addressing this, but it cannot fully account for the "reflexivity" of markets-where the actions of the algorithms themselves cause the very price moves they are trying to protect against.

The future scope of this research lies in the integration of Artificial Intelligence and Reinforcement Learning (RL) into market-making frameworks. As hinted at by Drissi (2022) and Ceffer et al. (2018), RL agents could theoretically discover more complex, non-linear strategies for inventory management and signal detection that go beyond the closed-form approximations currently in use. However, this also introduces the risk of "black box" behavior, where an algorithm might learn to engage in predatory practices or collusive behavior without explicit programming.

In the crypto-native space, the future of liquidity provision will likely be defined by "Automated Market Makers" (AMMs) and liquidity pools. The work of Kale (2025) suggests that the boundary between traditional FX hedging and crypto-asset management is blurring. As traditional institutions move into the digital asset space, they will bring their stochastic control models with them, but they must adapt to the "constant product" formulas and decentralized governance of the blockchain. The discussion must also consider the regulatory response; as the "last look" and internalization practices come under more scrutiny, we may see a push for more centralized, transparent clearing even in OTC-heavy markets like FX.

CONCLUSION

This research has synthesized a broad array of literature to provide a comprehensive view of the current state of algorithmic market making. From the foundational concepts of information asymmetry and the bid-ask spread to the cutting-edge mathematics of multi-asset internalization and dimensionality reduction, it is clear that the role of the market maker has become one of the most technologically and mathematically demanding positions in the financial industry.

The findings confirm that while algorithmic liquidity provision has generally led to narrower spreads and more efficient markets on average, it has also introduced new forms of systemic risk related to speed, internalization, and algorithmic correlation. The "liquidity mirage" remains a potent threat to market stability, particularly when the incentives of high-frequency intermediaries are not aligned with long-term market health. The complexity of managing inventory across multiple assets and venues requires sophisticated stochastic control mechanisms that can operate under significant model uncertainty and latency constraints.

In conclusion, the evolution of market making is a testament to the power of quantitative finance to transform economic structures. However, as the "arms race" continues and the "curse of dimensionality" is challenged by new approximation techniques, regulators and participants alike must remain vigilant. The goal should be to foster a market environment where the benefits of technological efficiency are balanced with the need for transparency, diversity, and robust price discovery. Future research should continue to explore the intersection of machine learning and market microstructure to ensure that the next generation of liquidity provision is not only faster but also more resilient.

REFERENCES

1. Aït-Sahalia, Y., Laeven, R. J., & Pelizzon, L. (2013). High Frequency Traders: Taking Advantage of Speed. Cambridge, MA.
2. Andersen, T., Bondarenko, O., & Gonzalez-Perez, M. T. (2021). A descriptive study of high-frequency trade and quote option data. Journal of Financial Economics.

3. Arumugam, D., & Varne, P. (2021). A game of hide-and-seek between proprietary and buy-side algorithmic traders: causal links with market quality. Applied Economics.
4. Bali, T. G., & Engle, R. F. (2006). Is there a risk–return trade-off? Evidence from high-frequency data. Journal of Applied Econometrics.
5. Benos, E., Brugler, J., Hjalmarsson, E., & Zikes, F. (2017). Interactions among high-frequency traders. Journal of Financial and Quantitative Analysis.
6. Bergault, P., Evangelista, D., Guéant, O., & Vieira, D. (2021). Closed-form approximations in multi-asset market making. Applied Mathematical Finance, 28(2), 101–142.
7. Bergault, P., & Guéant, O. (2021). Size matters for OTC market makers: General results and dimensionality reduction techniques. Mathematical Finance, 31(1), 279–322.
8. Berkowitz, S. A., Logue, D. E., & Noser, E. A. (1988). The total cost of transactions on the NYSE. Journal of Finance.
9. Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: frequent batch auctions as a market design response. Quarterly Journal of Economics.
10. Butz, M., & Oomen, R. (2019). Internalisation by electronic FX spot dealers. Quantitative Finance, 19(1), 35–56.
11. Capponi, A., Figueroa-López, J., & Yu, C. (2021). Market making with stochastic liquidity demand: Simultaneous order arrival and price change forecasts. arXiv preprint arXiv:2101.03086.
12. Cartea, Á., Donnelly, R., & Jaimungal, S. (2017). Algorithmic trading with model uncertainty. SIAM Journal on Financial Mathematics, 8(1), 635–671.
13. Cartea, Á., Jaimungal, S., & Penalva, J. (2015). Algorithmic and high-frequency trading. Cambridge University Press.
14. Cartea, Á., Jaimungal, S., & Ricci, J. (2014). Buy low, sell high: A high frequency trading perspective. SIAM Journal on Financial Mathematics, 5(1), 415–444.
15. Cartea, Á., Jaimungal, S., & Walton, J. (2019). Foreign exchange markets with last look. Mathematics and Financial Economics, 13(1), 1–30.
16. Ceffer, A., & Szabó, B. (2018). Trading by estimating the quantized forward distribution. Applied Economics.
17. Copeland, T., & Galai, D. (1983). Information effects on the bid-ask spread. The Journal of Finance, 38(5), 1457–1469.
18. Drissi, F. (2022). Algorithmic portfolio trading with signals and learning. Working Paper. SIAM Journal on Financial Mathematics, 13(1).
19. Hu, Y., et al. (2019). Information asymmetry, cluster trading, and market efficiency: evidence from the Chinese stock market. Economic Modelling.
20. Kale, A. (2025). FX Hedging Algorithms for Crypto-Native Companies. International Journal of Advanced Artificial Intelligence Research, 2(10), 09-14. https://doi.org/10.55640/ijaair-v02i10-02
21. Lai, H. C., et al. (2014). Relationship between the trading behavior of three institutional investors and Taiwan stock index futures returns. Economic Modelling.