Queuing Uncertainty

(a job market paper)

Bart Zhou Yueshen

current version: November 20, 2013
(first version: November 1, 2013)
Queuing Uncertainty

Abstract

In a high-speed trading environment, traders simultaneously react to public information not knowing the sequence in which their orders arrive at the exchange. A theoretical model is developed to capture such queuing uncertainty. Market makers strategically choose the size of their limit orders to compete for a common profit opportunity in liquidity provision. In equilibrium, liquidity overshoots—orders at the end of the queue make expected losses. Once realized, liquidity provision in the bottom of the queue is withdrawn, resulting in “flickering orders”. A boost in the trading speed amplifies the overshoot but the effect on order book dynamics (strategic order submission and cancellation) depends on the source of the speed increase. These predictions echo empirical evidence on and policy concerns over “quote stuffing”, order-to-trade ratios, and minimum quote life. The model points to an optimal level of queuing uncertainty, to which the exchange can steer by carefully randomizing the limit order queues.

Keywords: market microstructure, limit order market, queue, book depth, high frequency trading, flickering orders, speed

JEL code: D40, D43, D47, G10, G18, G19, L10, L13

(There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.)
1 Introduction

Since the automation of securities trading in the late 1990s, speed has become a salient feature of modern financial markets. The rise of the machines—technological advancement marked by the upsurge of algorithmic and high-frequency trading—has remarkably reduced trading latency from minutes (Biais, Hillion, and Spatt, 1995) to milliseconds (Hasbrouck and Saar, 2013), with the latest record in nanoseconds (Gai, Yao, and Ye, 2013).

As the trading speed approaches the limit imposed by nature\(^1\), it is arguable whether market participants can still perfectly condition their decisions on real-time market status. For example, when a trader sees an update of the limit order book, is this the true book status at that moment? Probably not: Among many potential delays, it takes time for the electronic message to travel from the exchange to the trader, and, during this period, the order book might evolve.

Even understanding the market condition \textit{ex post} is difficult: It took months for the U.S. Securities Exchange Commission (SEC) to publish a final report on the Flash Crash to explain what happened during the historic 20 minutes in which the entire U.S. equity markets experienced a 9 percent intra-day swing, down and then up.\(^2\) In Europe, MiFID II has expressed specific concerns over the impact of new technology on the quality of financial markets. The U.K. Foresight Project also highlighted the importance of understanding the new, high-speed trading environment.

Under the premise that traders \textit{cannot} perfectly condition on the true market status in real-time (especially in an extremely low-latency trading environment), this paper formally addresses the following questions: What does such imperfect conditioning imply for market participants’ optimal decisions? How can market quality, such as liquidity and allocative efficiency, be evaluated under such imperfection? How should the market be organized and what should regulators do, if anything, to deal with the consequences?

The main contributions are threefold. First, this paper develops a novel theoretical framework to analyze traders’ optimal decisions, accounting explicitly for their imperfect conditioning on the market status. Second, the model yields rich testable predictions, echoing evidence from the existing literature, on equilibrium

\(^{1}\) A photon travels in a vacuum from New York City to Chicago in about five milliseconds, an eternity in a trading environment with latency at nanoseconds.

\(^{2}\) The SEC Chairman Mary Schapiro emphasized “…that the regulator of the largest capital markets in the world cannot easily reconstruct trading … is not acceptable…” See Jonathan Spicer, Herbert Lash, and Sarah N. Lynch (2012, January 12),“Insight: SEC tightens leash on exchanges post ‘flash crash’”, \textit{Reuters}. 
Figure 1: **Game structure comparison.** This figure illustrates the difference between the conventional view and the new view (this paper) for low-latency trading. In panel (a), agents arrive sequentially and react to what previous agents have done. Panel (b) adds queuing uncertainty to panel (a). Nature queues the two agents randomly, but the two agents do not know which queue positions. The agents in panel (b) effectively play a simultaneous game.

liquidity provision and dynamics, e.g. “flickering depth”, “quote stuffing”, heavy electronic message traffic (in absence of fundamental shocks), and “holes” in limit order books. A competition channel is identified through which different types of trading latencies affect strategic liquidity provision differently. Finally, market quality is examined through a welfare criterion embedded in the model. The generated policy implications anchor recent regulatory debates regarding high-frequency trading, e.g. minimum order resting time, capping order-to-trade ratio, batching orders and randomizing queues, and etc. It is shown that trading efficiency can be improved by adjusting—sometimes increasing—the uncertainty in the market.

The theoretical contribution lies in the novelty of the model’s game structure. Figure 1 illustrates the idea. Consider a very short time period immediate after an information event. Panel (a) presents the extensive form of the classical treatment of such a scenario, where the sequential nature of the game guarantees that each agent can perfectly condition on the latest market status by observing what previous agents have done (for example, agent 2 observes the decision of agent 1). Panel (b) contrasts this classical view with “queuing uncertainty”, a specific form of the market imperfection discussed above. Nature queues the agents *randomly*, and the agents do *not* observe their queue position (note the information sets shown by the dashed
The model extends the seminal work of Glosten (1994) by incorporating the friction of queuing uncertainty. 
Glosten (1994) characterizes a “stable” equilibrium limit order book where at each price level, the marginal (last) limit order breaks even, earning zero profit in expectation. However, with queuing uncertainty, the equilibrium liquidity provision strictly exceeds the prediction of Glosten (1994): The equilibrium queues lengthen. Instead of breaking even, the marginal (last) limit order at each price level always makes expected losses, a phenomenon referred to as “liquidity overshoot”. The reason is that the new equilibrium condition becomes that at each price level, traders’ marginal orders break even individually in expectation, where this expectation is taken over all possible queue positions. With queuing uncertainty, an order in the top of the queue earns positive profit but a bottom order loses (as execution probability lowers and adverse-selection cost increases along the queue). For any queue realizations, there must be one trader whose order ends in the bottom of the queue and therefore the bottom order loses instead of breaking even: liquidity overshoot. Note that such “overshoot” is gauged only among the liquidity providers. Welfare implications, accounting also for liquidity demanders, are discussed below.

Effectively, the simultaneity puts the traders in competition, as each trader’s order size negatively affects the expected profitability of all others’. However, the queuing uncertainty weakens the strategic substitution effects and, consequently, amplifies the competition among the traders, leading to liquidity overshoot. To this extent, queuing uncertainty governs the level of competition.

Such liquidity overshoot is transitory and “unstable”. As some liquidity providers eventually realize that their orders are queued at the bottom and losing, the excess part will be canceled (if yet not executed). In a dynamic extension with repeated order modification, the model shows that such a revision option exacerbates the short-run overshoot: The liquidity providers compete more recklessly by submitting larger orders because they will be able later to cancel (revise) the orders that “unfortunately” sit in the bottom of the queue.

The dynamics of the displayed order book is described by the agents’ equilibrium strategies: There is intense overshoot of book depth in the short-run and the it gradually dies out in the long-run. The process

---

3 Model-wise, the game described in panel (b) is simultaneous, though it should be emphasized that the agents are not required to move simultaneously in the real world. Simultaneity is a modeling technique to reflect the combined information set.

4 The decreasing marginal profitability along the queue inherits from Glosten (1994). See also Hollifield, Miller, and Sandás (2004), Hasbrouck and Saar (2009), Liu (2009), and Raman and Yadav (2013) for empirical evidence.
complements Glosten (1994) by showing how the “stable” level of book depth is achieved over time. The model further reveals that in absence of shocks/news about the asset’s fundamentals, trading latency drops (faster trading speed) amplifies the liquidity overshoot in the short-run due to better revision options, and it takes many more rounds of revisions for the book depth to revert to its “stable” level. To this extent, the model explains how trading latency drops can lead to an increased cancellation/execution ratio (e.g. Gai, Yao, and Ye, 2013) and argues that the so-called “quote stuffing” might be an equilibrium outcome resulting from traders’ in ability to perfectly condition on real-time market conditions. This point is inspired by and similar to Baruch and Glosten (2013) who describe a mix-strategy equilibrium in a continuous price order book. The model framework can be readily used to formally analyze regulations concerning capping trade-to-quote ratios (implemented on, e.g., London Stock Exchange, Eurex Germany, Borsa Italiana among others) and minimum order resting time and the reasons why such policies may dampen liquidity provision.

However, latency drops can have ambiguous effects on the stabilization time of the order book. The model identifies two types of latencies. On the one hand, when agents’ reaction latency reduces, they are able to send more messages per unit of time, and hence, processing all orders takes longer time. On the other hand, when transmission latency reduces, the round-trip time of each order reduces, and the order book updates more frequently. All else being equal, the first type of latency reduction (for example, trading algorithm improvement and CPU upgrades) slows down the stabilization process, while the second (exchange server or optical fiber/microwave technology upgrades) hastens the convergence.

Although transitory, liquidity overshoot is nevertheless very important because it typically occurs immediately after an information event (a news announcement or simply an informed trade) and is followed by market orders trading on such information. Liquidity demanders (fundamental investors) benefit from the overshoot as they can meet their needs at lower cost. Such welfare concern echoes recent debate on how to regulate trading speed. Larry Harris suggests that “[r]egulatory authorities could require that all exchanges delay the processing of every posting, canceling and taking instruction they receive by a random period of between 0 and 10 milliseconds.”5 Correspondingly, EBS, a leading foreign exchange trading platform, recently implemented queue randomization: “[M]essages transmitting orders...will be bundled into batches and then run through a process that randomizes their place in the queue” and the randomization takes “between one and three milliseconds”.6

---

As a model application, this paper formally examines the above proposal, especially how queue randomization affects market quality. Intuitively, welfare generally improves with market liquidity, suggesting that adding queuing uncertainty to the market might improve efficiency because elevated competition will amplify the liquidity overshoot. However, there are cases where welfare is inverse U-shaped in the aggregate liquidity supply: Too much liquidity provision might indulge inefficient risk transfer. By properly randomizing the queues—essentially adjusting the degree of competition among liquidity providers—the exchange can steer the aggregate liquidity supply toward the socially optimal level. However, such randomization should be carried out with care because 1) the effectiveness of randomizing queues is bounded by the ex ante speed heterogeneity of all traders (and can be very limited); 2) the optimal randomness in queuing is a delicate measure and can be hard to determine in real-time trading; and 3) by adding queuing uncertainty, the exchange essentially deprives fast traders of their market power, disincentivizing participation in the long-run.

Additional related literature is discussed in section 2. A baseline model is developed in section 3. Section 4 explores the time dimension extensions and analyzes the book depth dynamics. Section 5 generates implications on allocative efficiency and market design. On the price dimension, an analysis of liquidity overshoot deep in the limit order book is provided in section 6. Section 7 then concludes. Appendix A provides a tractable micro-foundation for queuing uncertainty based on speed heterogeneity. Appendix B summarizes the notations. All proofs are collated in appendix C.

## 2 Related literature


---

7 Consider a risk-averse investor facing a sector of risk-neutral (or less risk-averse) market makers. Suppose that the book is very deep and the investor receives a private, imperfect signal that the asset value is very high. In this case, the investor buys all the asset offered in the market and yet, from a social planner’s point of view, this purely information-driven trade is inefficient because 1) the investor’s information rent offsets market makers’ adverse-selection cost and 2) risky asset positions are transferred from the less risk-averse market making sector to the risk-averse investor.
models can be found in Parlour and Seppi (2008).

The issue of uncertain sequentiality in limit order market previously focuses on the market orders and their execution. For example, Foucault, Kozhan, and Tham (2013) study “toxic arbitrage” activities in a foreign exchange market setting where triangular arbitrage opportunities are exploited by fast traders, hurting dealers who have uncertainty revising their quotes timely. In a related setting, Budish, Cramton, and Shim (2013) argue that the continuous time market design of financial asset trading is flawed in that the market makers’ inability to synchronize stale quotes with innovations in a reference asset results in a prisoners’ dilemma of technology arm race. Kervel (2013) studies liquidity provision by high-frequency traders who duplicate limit orders across venues and cancel immediately after trades occur in one of the venues. Hence, market order traders may or may not consume the visible liquidity. To compare, the current paper features the queue of limit orders, while the market order arrival is modeled to close the model.

Queuing uncertainty is also similar to the “contact-order uncertainty” in opaque over-the-counter (OTC) markets, as modeled in Zhu (2012). Despite the difference in the market setting (limit order book v.s. OTC market), in both models, liquidity providers do not know their queue positions. In the OTC setting, dealers compete strategically on their price quotes. In the limit order book setting, as the price grid is given by the book, the market makers compete on the size of their orders. The depth of the order book is, therefore, the focal liquidity measure of the current paper.

The model’s prediction on book depth dynamics recalls Baruch and Glosten (2013), who focus on the dynamics of quote prices in a limit order book with zero tick size, due to which liquidity providers can always undercut each other, resulting in a mixed-strategy equilibrium with “flickering quotes” (see Hasbrouck (2013) for an empirical analysis). Under the mixed-strategy equilibrium of their model, consistent with the current paper, agents’ decisions cannot be perfectly conditioned on by one another.\(^8\) Instead of quote prices, the current paper features short-run depth overshoot followed by immediate cancellation. Such “flickering depth” prediction agrees with Baruch and Glosten (2013) in that, the so-called “quote-stuffing” behavior by high-frequency traders, rather than gaming the trading system, might in fact be an equilibrium pattern as liquidity providers cancel their losing orders, the occurrence of which is nothing but an “unfortunate” real-

\(^8\) The drivers, however, are different: It is the zero tick size assumption in Baruch and Glosten (2013), while in the current paper it is the queuing uncertainty resulting from the non-zero latency at which any trader can register new market updates. The current paper, hence, does not have direct implication on quote prices, while Baruch and Glosten (2013) do not generate predictions on order book depth (rigorously speaking, depth is not properly defined in a zero tick size limit order market).
ization of the orders’ queue position. In particular, this paper shows that, given sufficient time for revision, the limit order book converges to the stable equilibrium described in Glosten (1994) (in contrast to Back and Baruch (2013), such convergence does not require infinite market makers).

Liquidity overshoot is evidenced in Sandås (2001), who shows that the break-even condition of Glosten (1994) does not hold on average and, instead, the marginal limit order loses. Such losses are attributed to the **negative** order processing cost (see tables 5 and 6 in Sandås, 2001), which is interpreted as a result of traders’ heterogeneous valuations for the asset. Queuing uncertainty provides an alternative equilibrium explanation that does not require heterogeneity in traders.

The current paper also adds to the debate on the impact of trading speed, e.g. due to the high-frequency traders, on market quality. See, among others, Pagnotta and Philippon (2012), Biais, Foucault, and Moinas (2013), Hoffmann (2013), and Menkveld and Zoican (2013) for related studies. Notably, the current model examines market quality only in terms of the gains from trade (competition on the intensive margin) and does not rely on the technology arm race (competition on the extensive margin), which is potentially socially costly. Accompanying the theory literature, empirical evidence on how low-latency trading technology affects market quality is also growing. See Hendershott, Jones, and Menkveld (2011) on the effect of the autoquote system on NYSE market quality; Jovanovic and Menkveld (2012) on the competition between an entrant platform, Chi-X, and the incumbent Euronext in Belgian and Dutch equity markets; Riordan and Storkenmaier (2012) on the speed improvement of Deutsche Boerse in 2007; Chaboud et al. (2013) on the impact of high-frequency trading on foreign exchange markets; and Hasbrouck and Saar (2013) on low-latency trading strategy in NASDAQ.

The model also emphasizes the important difference of the effects of latency drops on market quality. For example, while Hasbrouck and Saar (2013) find order book tends to be deeper when reaction latency is low (large amount of low-latency activity), studies on exchange speed upgrades—Riordan and Storkenmaier (2012) and Gai, Yao, and Ye (2013), for example—find lower transmission latency is associated with reduced displayed book depth. On colocation, Brogaard et al. (2013) and Frino, Mollica, and Webb (2013) find that upgrades and introduction both deepen the order book. The distinction in the types of latency drops is unclear in the above literature, as, for example, traders’ reaction latency probably correlate with the exchange’s transmission latency. Empirical methods to disentangle the effects are expected.

Outside the literature of financial market structure, queuing uncertainty closely relates to competition
in R&D. The “sprint race” feature of such games implies that the winner, who first accomplishes a technological innovation, takes all market share (as he becomes a monopolist protected by patent law) while the other competitors get nothing. The strategic competition in terms of research expenditure is reminiscent of the strategic choice of limit order size in the current paper. See, for example, Loury (1979), Lee and Wilde (1980), Reinganum (1981), and Grossman and Shapiro (1987) among many others. A notable difference in the game is that the payoffs to firms in the R&D literature exhibit a jump between the first and the second in queue as the winner takes all, while the payoffs along a limit order queue are smooth as the probability of execution and the cost of adverse selection are naturally continuous functions.

3 Baseline model

This section presents a fairly simple baseline model to illustrate the driver result of this study. The focus is on the liquidity provision on one price level in a very short period, immediately following some public information event. Such information event could be macro news, earnings announcement, or simply an update in the order book. More concretely, for example, when a sell market order hits the best bid price, the updated asset value typically drops and market makers might want to add ask orders at the best ask price. The sell market order is the information event and the best ask price is the battlefield analyzed in this section. The one-price focus is motivated by the empirical fact that there are few market orders that bite deep into the book, and most of the order activity concentrates on the top of the limit order book. The dynamic extension is studied in section 4 and the extension to multiple price levels is considered in section 6.

3.1 Model setup

The model is consistent with Glosten (1994). The new element is queuing uncertainty, as explained below. This baseline model studies only the liquidity supply side, leaving the demand side exogenous. Sections 5 endogenizes liquidity demanders.

Agents. There are $n$ risk-neutral agents indexed by $i \in \mathcal{N} := \{1, \ldots, n\}$, where $n$ is a positive integer (the special case of a continuum of agents is considered in section 4). These agents represent the market’s liquidity supply side, which is the focus of this section, and they will be referred to as market makers (to be consistent with the literature), though in reality these agents can be any traders who want to use limit orders.
at such occasions (e.g. endogenous liquidity providers, or ELPs, as studied in Anand and Venkataraman, 2013).

**Timing.** The baseline model is a one-period simultaneous game. After an information event, each market maker “simultaneously” and independently chooses his limit order size, \(q_i \geq 0\). These orders are then queued by nature according to a known distribution (see “queuing uncertainty” below; appendix A provides a tractable example for this distribution). Finally, a market order arrives (exogenously), trades occur, and market makers consume. The extensive form of the game with \(n = 2\) market makers is illustrated in panel (b) of figure 1.

**Limit order profitability.** The limit orders are perfectly divisible. Let \(\hat{\pi}(y)\) be the (expected) marginal profit of the \(y\)-th limit order at the given price level. Then the (expected) profit of the first limit order of size \(y \geq 0\) at the given price is \(\pi(y) = \int_0^y \hat{\pi}(x)dx\). For example, if a limit order of size 10 is added to a price level with existing depth of 5 units, it earns \(\pi(10 + 5) - \pi(5)\) in expectation.

**Assumption 1** (Quasi-top-of-queue advantage). The (expected) profit \(\pi(y)\) is twice-differentiable and is quasi-concave on \(y \in [0, \infty)\). Define its first order derivative to be \(\hat{\pi}(y)\). The two terminal conditions hold: 1) \(\hat{\pi}(0) > 0\) and 2) \(\lim_{y \to \infty} \hat{\pi}(y) < 0\). Finally, normalize \(\pi(0) = 0\).

For this section and the next, such a \(\pi(y)\) is assumed to be exogenously given. In section 5 (the welfare analysis), \(\pi(y)\) will be endogenized and it will be shown that assumption 1 holds exactly. Some remarks are in order.

**Remark 1.** Quasi-concavity implies the marginal profit, \(\hat{\pi}(y)\), is, loosely speaking, first positive and then negative (but never positive again) as \(y\) increases. That is, if one thinks of a limit order book as a set of queues of orders (at various prices), orders on top of each queue earn profit while those in the bottom lose, hence the term “(quasi-)top-of-queue advantage”. If \(\pi(y)\) is strengthened to be concave, the top-of-queue advantage is exact: The marginal profitability always decreases along the queue.

**Remark 2.** A quasi-concave \(\pi(y)\) is reasonable because orders in the bottom of the queue have lower execution probability and are subject to higher adverse-selection costs (large market orders tend to carry

---

This paper adopts Newton’s notation, variables with dot(s) overhead, to indicate derivatives.
strong signals). See the literature listed in footnote 4 for empirical evidence for decreasing marginal profitability.

Remark 3. The two terminal conditions are assumed to rule out triviality. First, \( \pi(0) > 0 \) guarantees participation: Otherwise, by quasi-concavity, \( \pi(y) \leq 0 \) for all \( y \geq 0 \) and each limit order at the price always loses. Second, \( \pi(\infty) < 0 \) rules out the uninteresting case where the marginal profit is always strictly positive (all market makers submit infinitely large limit orders).

**Queuing uncertainty.** Let \( \mathcal{K} \) be the collection of all possible queues.\(^{10}\) Ties (for example, both market makers 1 and 2 are at the same position in queue) are ruled out. Each queue is represented by an \( n \)-by-1 vector \( k \in \mathbb{N}^n \), where the \( i \)-th value \( k_i \in \mathbb{N} \) is the queue position of market maker \( i \). For example, if \( n = 3 \), a queue \( k = [2, 3, 1]^\top \) means market maker 1 is the second in the queue, market maker 2 the third, and market maker 3 the first. Fix a probability measure that defines queue distribution \( \mathbb{P}(K = k) \) for all \( k \in \mathcal{K} \). The distribution of the random vector \( K \) is known to all market makers.\(^{11}\)

### 3.2 Equilibrium

Consider market maker \( i \). He chooses his order size \( q_i \), given all others’ order sizes, to maximize his expected profit. He cares about the queue position of his order because of the top-of-queue advantage (assumption 1). For a given queue \( k \), write the aggregate size of the orders that queue before market maker \( i \)’s order by

\[
Q^<_{i}(k) = \sum_{j \in \mathcal{N}} q_j 1\{k_j < k_i\},
\]

where the superscript “\(<\)” emphasizes that \( Q^<_{i} \) excludes market maker \( i \)’s own order. A (pure-strategy) Nash equilibrium is a set \( q = \{q_1, \ldots, q_n\} \) such that \( \forall i \in \mathcal{N} \), given \( q_{-i} = q \setminus \{q_i\} \),

\[
q_i \in \arg \max_{q_i} \mathbb{E}\left[ \pi \left( Q^<_{i}(K) + q_i \right) - \pi \left( Q^<_{i}(K) \right) \right].
\]

The existence of such a Nash equilibrium is established by the following lemma.

---

\(^{10}\) \( \mathcal{K} \) is a finite set because there are at most \( n! \) possible queues.

\(^{11}\) Note that the specification allows speed heterogeneity: Some market makers can be faster than others. For example, in a two-agent game, market-maker 1 is said to be faster than market-maker 2 if \( \mathbb{P}(K = [1, 2]^\top) > \mathbb{P}(K = [2, 1]^\top) \). Appendix A characterizes such speed heterogeneity in terms of stochastic dominance.
Lemma 1 (Existence of Nash equilibrium). There exists a (pure-strategy) Nash equilibrium \( q \in [0, y^\equiv]^n \) where \( y^\equiv > 0 \) is the unique solution to \( \dot{\pi}(y) = 0 \) and at least one \( q_i \) is strictly positive.

Lemma 1 gives an upper bound, \( y^\equiv \), for each individual’s equilibrium limit order size. The intuition is as follows: Note that \( y^\equiv \) is the break-even point at which the marginal unit of limit order earns zero profit in expectation: \( \dot{\pi}(y^\equiv) = 0; \) the superscript “\( ^\equiv \)” emphasizes that the quantity breaks even. Therefore, no market maker will post a limit order exceeding \( y^\equiv \) units because the part beyond the break-even point always loses (by quasi-concavity of \( \pi(\cdot) \)), regardless of the order’s queue position.

Clearly neither \( q_i = 0 \) for all \( i \in \mathcal{N} \) or \( q_i = y^\equiv \) for all \( i \in \mathcal{N} \) \((n \geq 2)\) is an equilibrium. At least one market maker’s order size is an interior solution, and the first-order condition holds for him:

\[
\mathbb{E}_{\pi} \left( Q_i^\equiv(K) + q_i \right) = 0. \tag{2}
\]

This first-order condition implies the following proposition.

Proposition 1 (Liquidity overshoot). In equilibrium, liquidity overshoots in the sense that the last unit of the limit order earns negative expected profit, \( \pi(\sum_i q_i) \leq 0 \). The equality holds if and only if there is one market maker who is almost surely the first in the queue, i.e. \( \exists i \in \mathcal{N} \) such that \( \mathbb{P}(K_i = 1) = 1 \).

The intuition of this proposition is sketched in figure 2 with \( n = 2 \) market makers, whose orders are equally likely to be the first or the second in the queue. The downward sloping curve is the marginal profit \( \dot{\pi}(\cdot) \). It crosses the horizontal axis at \( y^\equiv \), which maximizes the profit \( \pi(\cdot) \). If there were only one market maker (who almost surely arrives first in queue), this \( y^\equiv \) would be the optimal quantity he would choose (c.f. the break-even condition of Glosten (1994), proposition 2). The two market makers, however, choose a symmetric equilibrium strategy \( q^\ast \) such that each market maker has a zero expectation, taken over all possible queue realizations, for his marginal profit. The first-order condition (2) simplifies to \( \frac{1}{2}\dot{\pi}(q^\ast) + \frac{1}{2}\dot{\pi}(2q^\ast) = 0 \) in a symmetric equilibrium. That is, they choose \( q^\ast \) such that the absolute value of \( \dot{\pi}(q^\ast) \) and \( \dot{\pi}(2q^\ast) \) are the same. Clearly, by quasi-concavity of \( \pi(\cdot) \), the marginal profit of the very last unit order in the book must be negative. In aggregate, therefore, liquidity overshoots.

The shaded area in figure 2 shows the size of the overshoot inefficiency. The key friction that creates such inefficiency is queuing uncertainty. If the queue is deterministic (panel (a) of figure 1), then the first market maker in the queue will submit an order of size \( y^\equiv \) and the maximum aggregate profit is achieved. It
Figure 2: Equilibrium illustration with $n = 2$ market makers. This graph illustrates the optimal supply decisions with two market makers, each having probability half of being the first in the queue. The downward sloping curve is the marginal profit function $\dot{\pi}(q)$. It crosses $\dot{\pi}(q) = 0$ at $q = y^\equiv$. In a symmetric-strategy equilibrium, the first-order condition (2) requires that the expected marginal profit be zero. That is, $\frac{1}{2}\dot{\pi}(q^\ast) + \frac{1}{2}\dot{\pi}(2q^\ast) = 0$, which implies that $\dot{\pi}(q^\ast) = \dot{\pi}(2q^\ast)$. The shaded area represents the (expected) loss that the last-in-queue market maker suffers.

should be noted that such inefficiency, however, is not necessarily socially costly: The overshoot of liquidity might benefit the demand side, which so far remains exogenous. Welfare analysis is deferred until section 5.

To better understand the intuition of proposition 1, note that the game is reminiscent of oligopolistic competition among producers of a homogeneous good: The supply of one producer (market maker) negatively affects the profit of his competitors. Indeed, the liquidity provision in limit order market can be viewed as strategic substitutes, as defined in Bulow, Geanakoplos, and Klemperer (1985):

Proposition 2 (Strategic substitution). Suppose $\pi(\cdot)$ is concave on $[0, \sum_i q_i^\ast]$ where $\{q_i^\ast\}$ is an equilibrium with $n \geq 2$ market makers. Then the market makers’ limit orders are strategic substitutes and, further, the substitution rate is lower than unity in absolute value. Mathematically, $\partial^2E[\pi(Q_i^j + q_i)]/(\partial q_i \partial q_j) \leq 0$ and $\partial q_i^\ast/\partial q_j \in (-1, 0]$ for all $i, j \in \{1, 2, \ldots, n\}$ and $i \neq j$.

Remark. Note that quasi-concavity is strengthened to concavity to guarantee a monotone decreasing $\dot{\pi}(\cdot)$ on the relevant support, akin to a monotone decreasing demand function.

Strategic substitutes means whenever a market maker $i$ increases his supply, he “threatens” to capture the others’ profit (when he queues to the front). However, the queuing uncertainty bounds the substitution rate. This is because other market makers “underweight” such threats as there is always non-zero probability
for market maker \( i \)'s order to queue behind his competitor's, in which case the "threat" is ineffective. As such threats are "underweighed", the market makers engage in fierce competition which results in liquidity overshoot; that is, the last unit of the limit order no longer breaks even but loses. The key friction is queuing uncertainty, which leads to the intensified competition.

### 3.3 Example

The baseline model above can be solved in closed-form if one is willing 1) to approximate the profit function by a second-order Taylor expansion and 2) to assume that market makers are homogeneous—all queue realizations are equally likely and they use symmetric strategies. That is, assume that \( \pi(y) \approx -\frac{1}{2}y^2 + y^\pi y \), which gives \( \pi(y) = y^\pi - y \), for all \( y \geq 0 \) and that \( \mathbb{P}(K = k) = 1/(n!) \) for all \( k \in \mathcal{K} \).

The first-order condition (2), under symmetric strategy restriction, simplifies to

\[
\sum_{k=1}^{n} \frac{1}{n} \pi(kq) = \frac{1}{n} \sum_{k=1}^{n} (y^\pi - kq) = 0,
\]

and the unique symmetric, pure-strategy Nash equilibrium is that all market makers choose \( q^* = 2y^\pi/(n+1) \), and the resulting total book depth is \( nq^* = 2ny^\pi/(n+1) \). Note that though \( q^* \leq y^\pi \), the total depth weakly exceeds \( y^\pi \) (because \( 2n/(n+1) \geq 1 \)) and the equality holds only at \( n = 1 \).

Under this equilibrium, it can be computed that the strategic substitution effect is \( \frac{\partial q^*_i}{\partial q_j} = \mathbb{P}(K_j < K_i) = (n-1)/(2n) \in [0, 1/2) \), always less than unity.

### 4 Book depth dynamics

This section builds on the baseline model to analyze the book depth dynamics over time: The market makers are allowed to revise their existing (unexecuted) limit orders. As before, the focus remains on the book depth at a single price level in a short time interval immediately following some information event. A heuristic discussion is provided first to illustrate the idea.

**Order book dynamics: Overshoot followed by immediate cancellation.** Consider the following simplified scenario: Three market makers simultaneously submit limit orders to compete for a profitable \( \pi(y) \), satisfying assumption 1. Suppose the three orders are of the same size, \( q (> 0) \). For this example, let \( q < y^\pi < 3q \), that is, there is liquidity overshoot. After the submission, a random order queue is formed at
the exchange server and as the queue is processed, three book updates are sequentially disseminated to the market makers.

When the first update is observed, the displayed depth becomes $q$, while the “true” depth is $3q$. Seeing the update, one of the three market makers is “happy”: His order is queued in the top, earning positive expected profit ($\pi(q) > \pi(y) = 0$). However, the other two market makers become “worried” as they realize that their orders are now more likely to be in the bottom of the queue; that is, their marginal orders lose in expectation. Therefore, these market makers with unresolved queuing uncertainty have an incentive to revise down their order sizes to some $q' \in (0, q)$, which amounts to two additional revision orders submitted to the exchange. In equilibrium, there will still be liquidity overshoot after the revision: $q + 2q' > y$ (proved later in the formal analysis). The “happy” market maker has no incentive to submit new orders or modify his existing order, as shown below.

When the second update is observed, the displayed depth is $2q$, while the true depth is $q + 2q'$. A second market maker is “happy” when it turns out that his order is second-in-queue and after the previous revision, it is earning positive profit in expectation ($q + q' < y < q + 2q'$). The other market maker is disappointed because his order is in at bottom, part of it losing, and he wants to cancel the losing part. Therefore, another revision order is submitted: The last-in-queue market maker cancels the losing part of his order, $q + 2q' - y$.

As the second update of the book fully resolves the queuing uncertainty, no more further revisions occur. In total, six orders are submitted in this example: Three new orders followed by three immediate revision/cancellation orders. The displayed book depth first increases and then decreases:

$$0 \rightarrow q \rightarrow 2q \rightarrow 3q \rightarrow (q + 2q') \rightarrow y \quad \text{(book stabilized).}$$

Note that from $2q$ until $y$, there is liquidity overshoot.

The rest of this subsection sets up a tractable framework that builds on this example to analyze the magnitude of liquidity overshoot, the number of revision orders, and the stabilization process of the order book. Apart from the results illustrated above, the model investigates how (different types of) trading latencies

\[12\] In reality, inferring queue position can be tricky. After submission, limit order traders receive confirmation and book update separately from the exchange. The confirmation acknowledges with a time stamp that the exchange has processed the order. However, depending on the platform, the confirmation may (for example, in Eurex) or may not contain the order’s queue position. Even if there is no queue position in the confirmation, the trader can still (imperfectly) infer it by comparing the time stamps of the confirmation and the recent book updates. This section assumes perfect inference of queue position. The case of no inference degenerates to the analysis in section 3, where there is no order revision after submission. These institutional insights are obtained from very helpful discussions with Andrei Kirilenko and with Bernard Hosman.
affect the equilibrium dynamics.

4.1 Model setup

The setup follows the baseline model. The top-of-queue advantage assumption 1 needs to be slightly strengthened:

**Assumption 2** (Top-of-queue advantage, strengthened). Assumption 1 holds and, further, \( \pi(y) \) is concave on \([0, y^\infty)\) where \( \dot{\pi}(y^\infty) = 0 \).

This strengthened version of top-of-queue advantage is micro-founded by various commonly used frameworks in market microstructure literature, as shown in section 6 later. The next assumption simplifies the analysis by switching from an oligopolistic competition game to a perfect competition environment.

**Assumption 3.** There is a continuum of homogeneous market makers of mass 1.

*Remark 1.* The continuum of market makers has two implications. First, each market maker is of infinitesimally small market power (zero-measure mass) and therefore any individual’s strategy does not affect the aggregate depth. Second, because there are infinitely many market makers, the liquidity supply business is essentially in perfect competition: The aggregate expected profit is zero in equilibrium, as will be shown.

*Remark 2.* The homogeneity in market makers implies a specific distribution for each individual’s order queue position: \( K_i \) is i.i.d. uniform on \([0, 1]\) for all \( i \in [0, 1] \). It also allows the analysis to focus on symmetric equilibrium.

*Timing.* The time line is extended as below and is illustrated in figure 3. Time continuously runs from 0 to infinity. The game ends at time \( T \), the random arrival time of the market order. Suppose \( T \) has a c.d.f. of \( G(T) > 0 \) with a continuous support on \( T \geq 0 \).

*Two latencies.* The model distinguishes two different types of latencies. First, there is a gap \( \delta (> 0) \), referred to as the reaction latency, for each market maker to submit two consecutive orders (either addition or cancellation) to the exchange; that is, all market makers submit new orders only at time \( t \in \{0, \delta, 2\delta, \ldots \} \).
Second, there is transmission latency, \( \eta > 0 \), defined below. Let the random round-trip time of each submitted order be i.i.d. from the uniform distribution on the support \([0, \eta]\). Given the unit mass continuum of market makers (assumption 3), it follows that if they all submit a new order at time \( t \), at each moment between \( t \) and \( t + \eta \), a flow of book updates of mass \( dt/\eta \) will be disseminated to and observed by all market makers. To avoid complication of information asymmetry among market makers, each book update is assumed to arrive to all at the same time.\(^{13}\)

The reaction latency is motivated by the fact that it takes strictly positive amount of time—for example, CPU time—for each market maker to make a decision. Though admittedly the technology advancement has reduced the reaction latency to the level of nanoseconds, it has not reached zero (and neither will it). The transmission latency measures how soon, after an order is issued, the market participants can observe the update. It involves the delays of message travel time and the exchange server’s processing time. To sum up, market makers’ reaction latency determines the intensity at which orders are generated, while the transmission latency sets the rate at which the changes made by these orders are displayed.

To intuitively understand the difference between the two latencies, it may be useful to compare monitoring the order book to filming a movie. Figuratively, the reaction latency is like the frame rate (usually measured in frames per second, fps) of the movie: The lower is the latency, the more frames (order book changes) are generated each second. On the other hand, the transmission latency is like the “fast-forward”

---

\(^{13}\) The setup here is consistent with Abreu and Brunnermeier (2003) who assume that in each instant of time there is a cohort of mass \( 1/\eta \) agents who becomes aware of mispricing in the economy. They refer to this period \([t, t + \eta]\) as the “awareness window”.

---
button that governs how fast the film is played, or how fast the order book changes are displayed.

**Time priority.** Time priority is enforced as before. That is, new orders submitted at later rounds always move to the end of the existing queue. Modeled after the real-world practices, when the size of a limit order is revised down, its queue position is not affected; but when the size is revised up, the order moves to the end of the queue; for example, see *Conditions for Trading at Eurex Deutschland and Eurex Zürich.*

### 4.2 Equilibrium

The focus is restricted to pure, symmetric-strategy equilibrium. Suppose the unit mass of market makers all submit an order of size \(q(0)\) at \(t = 0\). Given the flow of book updates at rate \(1/\eta\), the displayed order book depth is (almost surely) \(t q(0)/\eta\) for all \(t \in [0, \eta]\). Then, by \(t = h\bar{\delta}, h \in \{0, 1, \ldots\}\) such that \(h\bar{\delta} \leq \eta\), a mass of \(h\bar{\delta}/\eta\) of the initial batch of orders’ queuing positions are revealed. Consequently, a mass of \(\min\{1, h\bar{\delta}/\eta\}\) market makers observe the queue positions of their initial orders, while the other \(\max\{0, 1 - h\bar{\delta}/\eta\}\) mass of market makers still face queuing uncertainty.

Denote the total rounds of order revisions by \(\bar{h}\); that is, for \(t \geq (\bar{h} + 1)\bar{\delta}\), no market makers seek to revise their orders any more. The following two results will be verified along the analysis below:

**Result.** 1) **There is at least one revision, i.e.** \(\bar{h} \geq 1\). 2) For each \(h \in \{0, \ldots, \bar{h} - 1\}\), only the market makers whose initial orders are still subject to queuing uncertainty revise; the others do nothing.

The intuition behind these two results follows the liquidity overshoot (which, as will be shown, persists throughout). For result 1), there are always some market makers who will find their orders end up in the bottom and make expected losses, and therefore, at least one round of cancellation is needed to revert the book depth to its stable level, \(y^=\). Similarly, for result 2), as soon as a market maker has his order’s queuing uncertainty resolved, he no longer wants to revise it further: If his order queues before the break-even level of \(y^=\), it earns positive expected profit and will stay; otherwise, the order is immediately fully canceled.

To this point, a competitive, symmetric, pure-strategy equilibrium can be concisely described by a path of \(\{q(h)\}_{h=0}^{\bar{h}-1}\), where each \(q(h)\) is the after-revision order size chosen by all active market makers, such that each market maker maximizes his *per capita* expected profit at each round \(h\). In particular, at \(h = 0\), there is no revision and \(q(0)\) refers to the initial order size by all market makers. The terminal condition of \(q(\bar{h})\) will be solved together with the value of \(\bar{h}\).
Fix a candidate equilibrium path of \(\{q(h)\}\). Recall that there is a mass of \(\delta/\eta\) market makers who resolve their queuing uncertainty each round. Then the true book depth—the depth computed by aggregating up all submitted orders—after the \(h\)-th revision becomes

\[
y(h) = \sum_{j=0}^{h-1} \int_{j \frac{\delta}{\eta}}^{(j+1) \frac{\delta}{\eta}} q_j(j) \, dj + \int_{h \frac{\delta}{\eta}}^{1} q_i(h) \, dj = \sum_{j=0}^{h-1} \frac{\delta}{\eta} q(j) + \left(1 - h \frac{\delta}{\eta}\right) q(h).
\]

(The second equality holds by symmetry.) Note that the true depth is decomposed into a stabilized component and a to-be-revised component. The first component represents the depth contributed by the inactive market makers whose queuing uncertainty has been resolved. The second component is the depth contributed by all active market makers who are still subject to queuing uncertainty and therefore, may revise in later rounds. The residual expected profit for the active market makers, given the stabilized component in the true depth (equation (3)), becomes

\[
\pi_h(y) = \pi \left( y + \sum_{i=0}^{h-1} \frac{\delta}{\eta} q(i) \right) - \pi \left( \sum_{i=0}^{h-1} \frac{\delta}{\eta} q(i) \right),
\]

which is a left-down parallel shift of \(\pi(y)\). It inherits from \(\pi(y)\) a break-even point \(y^\circ_h = y^\circ - \frac{\delta}{\eta} \sum_{i=0}^{h-1} q(i)\) and satisfies assumption 2.

Consider next the optimization problem of the \((1 - h\delta/\eta)\) mass of active market makers (whose initial orders are still subject to queuing uncertainty). To prepare for the Bellman equation, first construct an expression of their expected profit upon execution. By symmetry, each of these market makers knows that his own initial order will be queued at position \(K_h\), a uniformly distributed random variable on \([0, 1 - h\delta]\).

Upon execution, a market maker’s per capita expected profit of his order, given all others’ order sizes, is

\[
\mathbb{E} \left[ \pi_h \left( Q^\circ_i + q_i \Delta \right) - \pi_h \left( Q^\circ_i \right) \right]/\Delta,
\]

where \(\Delta\) is the mass of the market maker and \(Q^\circ_i\) is the random cumulative size of the orders queuing in front of market maker \(i\)’s order. Let all active market makers, other than \(i\), choose the same order size of \(q(h)\) (symmetric equilibrium). Hence, if market maker \(i\)’s initial order position turns out to be \(k_h \in [0, 1 - h\delta/\eta]\), then \(Q^\circ_i = \int_0^{k_h} q(h) \, dj = k_h q(h)\). Further, by the continuum assumption, \(\Delta\) is infinitesimally small and the per capita expected profit upon execution becomes, in the limit,

\[
\lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E} \left[ \pi_h \left( Q^\circ_i + q_i \Delta \right) - \pi_h \left( Q^\circ_i \right) \right] = \mathbb{E} \pi_h \left( Q^\circ_i \right) q_i = \mathbb{E} \pi_h \left( K_h q(h) \right) q_i.
\]
Using the above expression, the Bellman equation of market maker $i$ can be written as

$$u_h = \max_{q_i} \mathbb{P}(T \leq (h + 1)\delta) \left[ \mathbb{E}_{\pi_h}(K_h q(h)) q_i \right]$$

$$+ \mathbb{P}(T > (h + 1)\delta) \left[ \mathbb{P}\left(K_h \leq \min\left\{ k^*_h, \frac{\delta}{\eta} \right\} \right) \mathbb{E}\left(\pi_h (K_h q(h)) | K_h < \min\left\{ k^*_h, \frac{\delta}{\eta} \right\} \right) 
+ \mathbb{P}\left(K_h > \min\left\{ k^*_h, \frac{\delta}{\eta} \right\} \right) u_{h+1} \right].$$

The first term on the right-hand side is the expected profit if the market order arrives before the next revision. The second term—if the market order arrives very late, allowing next the revision—has two components, depending on whether the market maker’s initial order will be queued to the front or to the rear. The cut-off point $k^*_h$ determining “front” or “rear” is defined, for $h < \bar{h}$, as

$$k^*_h(q(h)) = \frac{1}{q(h)} \left( y^\pi - \sum_{j=0}^{h-1} \frac{\delta}{\eta} q(j) \right).$$

That is, if $K_h > k^*_h$, the order queuing beyond $y^\pi$ and should be canceled.

A significant simplification is that $q_i$ does not affect the continuation value $u_{h+1}$ for all $h \in \{0, 1, ..., \bar{h}\}$, thanks to the continuum of market makers, each of whom has infinitesimally small (zero) influence on the total depth. Mathematically, $\partial u_{h+1}/\partial q_i = 0$. Thus, the first-order condition simplifies to

$$\mathbb{P}(T \leq (h + 1)\delta | T > h\delta) \mathbb{E}_{\pi_h}(K_h q(h))$$

$$+ \mathbb{P}(T > (h + 1)\delta | T > h\delta) \mathbb{P}\left(K_h \leq \min\left\{ k^*_h, \frac{\delta}{\eta} \right\} \right) \mathbb{E}\left(\pi_h (K_h q(h)) | K_h < \min\left\{ k^*_h, \frac{\delta}{\eta} \right\} \right) = 0$$

and after evaluating the expectation terms (recall that $K_h$ is uniformly distributed) and the probability terms, the first-order condition becomes

$$\pi_h ((1 - h\delta)q(h)) + \frac{1}{\lambda(h; \delta)} \pi_h \left( \min\left\{ k^*_h(q(h)), \frac{\delta}{\eta} \right\} q(h) \right) = 0,$$

where $\lambda(h; \delta) := (1 - G((h + 1)\delta))/\left(G((h + 1)\delta) - G(h\delta)\right)$ is the inverse (discrete) hazard rate of $T$. The result is a system of difference equations which recursively updates $\pi_h(y)$ according to equation (4) and $k^*_h$ according to equation (5).\(^{14}\)

\(^{14}\) Note that the left-hand side of the first-order condition (6) (i.e. $du_h/dq_i$) does not depend on market maker $i$’s order size $q_i$. Hence, as soon as this first-order condition holds, market maker $i$ is indifferent to any order size: As an infinitesimally small agent in the continuum, he has zero-mass and no influence on the aggregate liquidity supply. To close the equilibrium, the symmetric strategy assumption kicks in so that market maker $i$ also chooses $q_i = q(h)$.  

19
Equation (6) is a zero-profit condition for all (active) market makers. To see this, consider the baseline model of no revision opportunity; that is, \( \lambda(h; \delta) \to \infty \), nullifying the second term in equation (6). Observe further that the resulting first order condition is \( \mathbb{E} \pi_h (K_h q(h)) = \pi_h ((1 - h\delta)q(h)) = 0 \), consistent with the baseline version as in equation (2). The only difference is that under the continuum assumption, the competition among market makers is perfect and hence they earn zero profit in aggregation.

The zero-profit condition (6) implies the following version of liquidity overshoot:

**Lemma 2 (Persistent liquidity overshoot).** There is liquidity overshoot at all rounds of order revision until the order book stabilizes.

The results 1) and 2) claimed at the beginning of the equilibrium analysis can now be verified using lemma 2. For point 1): Suppose \( \bar{h} = 0 \), that is, there is no revision after an initial round of submission. Then, the initial submission must stabilize the book at \( y(0) = 1 \cdot q(0) = \bar{y} \), contradicting the overshoot implication of equation (6). For point 2): The market maker with queued order at the \( h \)-th round can submit, if desired, a new order or revise his existing order. A newly submitted order will queue at the end, implying expected loss because there is already liquidity overshoot. Revising down the size of the existing order only reduces the expected profit, while revising up the size moves the order to the end of the queue, losing in expectation. Hence, there is no incentive for the market makers with resolved queuing uncertainty to do anything.

It remains to determine \( \bar{h} \), the last round revision. The recursion stops if there is no unresolved queuing uncertainty; that is, if the stabilized component in the true depth exceeds the break-even level: \( \sum_{h=0}^{\bar{h}} \delta q(h)/\eta \geq \bar{y} \). Then the last revision always leads to \( q(\bar{h}) = 0 \) because any orders queued beyond \( \bar{y} \) lose in expectation. Rearranging the inequality above using \( q(\bar{h}) = 0 \) gives the stopping criterion

\[
\frac{\delta}{\eta} \geq \frac{1}{q(\bar{h} - 1)} \left( \bar{y} - \sum_{h=0}^{\bar{h}-2} \frac{\delta}{\eta} q(h) \right) = k_{\bar{h}-1}^*.
\]

That is, as soon as \( k_{\bar{h}}^* \geq \delta/\eta \), the recursion stops in the next round, i.e. \( \bar{h} = h + 1 \). The following lemma summarizes the analysis above and gives a pure-strategy, symmetric equilibrium.

**Lemma 3 (Equilibrium order submission and revision).** At \( t = 0 \), all market makers submit a limit order of size \( q(0) \) where \( q(0) \) solves equation (6) with \( h = 0 \). At \( t = h\delta \) for all \( h \in \{1, ..., \bar{h} - 1\} \), if a market maker observes the queue position of his initial order to be \( k < h\delta/\eta \), he does nothing; otherwise, he submits a revision order that changes his order size to \( \min\{q(h - 1), q(h)\} \) where \( q(h) \) can be recursively
solved from equations (4), (5), and (6). The recursion stops at $\bar{h}$ where $\bar{h} = \min \{ h | h \geq 1, k_{h-1}^* \leq \delta/\eta \}$.

4.3 Latency and order book dynamics

This subsection focuses on the effects of latency reduction on order book dynamics: Do lower latencies dampen or amplify liquidity overshoot? What are the implications on the stabilization of the order book? How do the effects of reaction latency $\delta$ and of transmission latency $\eta$ differ?

As seen in the previous subsection, the generic functional form of $G(\cdot)$ suffices to characterize the equilibrium book depth dynamics. But to answer the above questions, the following parametrization is imposed on the arrival time of the market order.

Assumption 4 (Market order arrival). The market order arrival time $T$ follows exponential distribution with $\mathbb{E}T = \tau > 0$. That is, $G(T) = 1 - e^{-T/\tau}$ for all $T \geq 0$.

Remark 1. Exponential distribution is often used in modeling arrival time. In particular, the memorylessness property implies that the arrival probability in a fixed time interval is constant: $\mathbb{P}(T \leq (h + 1)\delta | T > h\delta) = 1 - e^{-\delta/\tau}$ for all $h \geq 0$. To this extent, the assumption is consistent with, for example, Foucault (1999) who assumes the trading process ends with the same probability in each period.

Remark 2. Note that assumption 4 implies $\partial G(T)/\partial \delta = \partial G(T)/\partial \eta = 0$. That is, changes in reaction or transmission latency does not affect the arrival probability of the market order. This would be the case if the market orders are submitted only by investors subject to shocks of information and liquidity needs, whose arrival processes are plausibly independent of trading technology changes. A more realistic setting should allow the market orders’ expected arrival time $\tau$ to be, for example, positively correlated with $\delta$. The positive correlation, however, hinders the tractability of the model and, yet, numerical procedures suggest that the results shown in this subsection remain qualitatively the same with or without a moderate correlation. For this reason, the independence of $\tau$ and the latencies $\delta$ and $\eta$ is maintained.

The magnitude of short-run liquidity overshoot depends on the initial limit order size $q(0)$. The equilibrium level of $q(0)$ should satisfy the zero-profit condition at $h = 0$:

$$\pi(q(0)) + \frac{1}{\lambda(\delta)} \pi \left( \min \left\{ y^*, \frac{\delta}{\eta} q(0) \right\} \right) = 0.$$
(Note that \( k^*_0(q(0)) = y^\gamma / q(0) \).) In this equation, \( q(0) \) is expressed as an implicit function of the latency parameters \( \delta \) and \( \eta \).

**Proposition 3 (Short-run liquidity overshoot and latency).** *Cateris paribus, short-run liquidity overshoot amplifies if either reaction latency or transmission latency reduces.* That is, \( \partial q(0) / \partial \delta \leq 0 \) and \( \partial q(0) / \partial \eta \leq 0 \).

The intuition behind the above result is based on the market makers’ improved order revision opportunity. With lower reaction latency, the market makers can react to the updates of market conditions at higher frequency. With lower transmission latency, the market makers observe more flow of information (better monitoring) in any fixed time interval. Therefore, the revision opportunity is better than in the high latency case, their expected profit is higher, and they will more fiercely compete and generate more pronounced liquidity overshoot.

Figure 4 illustrates the amplified short-run over supply. Recall that the initial batch of orders (unit mass) is fully displayed by time \( \eta \) (almost surely; see the definition of transmission latency). The peak at \( t = \eta \) increases as \( \delta \) decreases from 0.5 to 0.1 in panel (a) and as \( \eta \) decreases from 1.2 to 0.7 in panel (b). (The numerical procedure and the parametrization are described at the end of this subsection.)

The improved revision opportunity implies that there will be more order revisions before the order book stabilizes. This raises the question whether the order book stabilization process slows down as the latencies reduce. To investigate this question, first note that at each but the last revision, the amount of orders submitted is \( (1 - h\delta / \eta) \), which is the mass of market makers with unresolved queuing uncertainty. Second, the last batch of order amounts to mass \( (1 - (\bar{h} - 1)\delta / \eta) - k^*_\bar{h}-1 \), which is the mass of market makers who realize that their initial orders are queued behind \( y^\gamma \) (hence canceling fully). Finally, recall that the transmission latency is defined so that it (almost surely) takes \( \eta \) units of time to display each unit mass of orders. Then define the stabilization time of the displayed order book by

\[
t^\gamma(\delta, \eta) := \eta \cdot \left[ \sum_{h=0}^{\bar{h}-1} \left( 1 - h \frac{\delta}{\eta} \right) + \left( 1 - (\bar{h} - 1) \frac{\delta}{\eta} \right) - k^*_\bar{h}-1 \right] = (\bar{h} + 1) \left( \eta - \frac{1}{2} \delta \bar{h} \right) + \left( \delta - \eta k^*_\bar{h}-1 \right).
\]

Note that \( \bar{h} \) is implicitly defined by \( \delta \) and \( \eta \), and because \( \bar{h} \) can take only integer values, it turns out to be a step-function in the latency parameters. The direction of the step function can be signed:
(a) Varying reaction latency, $\delta$

(b) Varying transmission latency, $\eta$

Figure 4: **Order book depth dynamics with varying latency.** The left panel is the contour plot of displayed depth over time $t \in (0, 6)$ (horizontal axis) with varying reaction latency $\delta \in (0, 0.5)$ (vertical axis). The right panel plots the depth dynamics for four specific levels of $\delta \in \{0.5, 0.25, 0.15, 0.1\}$. The numerical procedure is performed under the parametrization of $\pi(y) = y - y^2$, $\eta = 1$, and $\tau = 1$.

**Lemma 4 (Number of revisions).** When the latencies are low, $\bar{h}$ is a right-continuous, decreasing step function in $\eta$ and is a left-continuous, increasing step function in $\delta$.

The lemma helps the derivation of the following proposition.

**Proposition 4 (Order book stabilization and latency).** In a low latency environment, cateris paribus, the stabilization of the order book shortens if transmission latency lowers but prolongs if reaction latency lowers. That is, $\bar{r}(\delta, \eta) > \bar{r}(\delta, \eta')$ and $\bar{r}(\delta, \eta) < \bar{r}(\delta', \eta)$ for $0 < \delta' < \delta$ and $0 < \eta' < \eta$.

**Remark.** Comparative statics with respect to the latencies are nontrivial as the net effect depends on the indirect effects from all revision order sizes. Signing the comparative statics in general seems difficult.

A key simplification to facilitate lemma 4 and proposition 4 results from the focus on low latency environment, in which $\delta$ and $\eta$ are small and the indirect effects are dominated by the direct effects. The details can be found in the proof.

The countervailing effects of transmission latency and reaction latency can be understood as follows: The former latency governs how fast the market events are played while the later measures how many such
events are generated every moment. As the transmission speed increases (\( \eta \) drops; the movie being fast-forwarded), the stabilization occurs sooner than before. As more market events are being generated every moment (\( \delta \) drops; more frames inserted to the movie), naturally, it takes longer to have all these events processed and disseminated. The pattern is illustrated in figure 4.

The above proposition might seem trivial, but it has important implications for empirical works. In particular, it emphasizes the difference between the two types of latencies. A technology improvement that reduces the exchange’s latency perhaps should be distinguished from an innovation in the traders’ trading technology. For example, consider an empiricist who obtains an order book dataset that covers a period during which both reaction and transmission latency are reduced. Intuitively, the two latencies should change in a correlated way. For example, when the exchange provides colocation service (reducing transmission latency), it attracts high-frequency traders who have ultra-low reaction latency. How does the speed upgrade affect the book depth as seen from the empirical examination?

The answer depends on both the dominant force from the latency reductions and the sampling frequency chosen by the empiricist. Consider figure 4. Fix an observation time \( t \) and note that as the reaction latency drops in panel (a), the observed book depth increases along the vertical slice. In panel (b), it can be seen that reducing the transmission does not monotonically affect the displayed book depth at a fixed observation time \( t \).

Indeed, empirical evidences on how latency drops affect order book depth do not seem to be unanimous. For example, on the transmission latency reductions, Riordan and Storkenmaier (2012) and Gai, Yao, and Ye (2013) find exchange speed upgrades lead to reduced order book depth; on the other hand, Frino, Mollica, and Webb (2013) and Brogaard et al. (2013) find introduction and upgrades of colocation service by exchange markets improve observed book depth. The current model provides a plausible explanation for the “disagreement”. It emphasizes that empirical works should construct depth (and other liquidity) variables with extra care when analyzing how latency affects market quality.

**Numerical illustration.** To illustrate the above comparative static results, the order book depth dynamics are plotted in figure 4 with varying \( \delta \) and \( \eta \), respectively, under the following parametrization: \( \pi(y) = y - y^2/2 \) (implying \( y^= = 1 \)) and \( \tau = 1 \). For panel (a), \( \eta \) is fixed at \( \eta = 1 \) and for panel (b), \( \delta \) is fixed at \( \delta = 0.15 \). The quadratic \( \pi(y) \) can be motivated as a coarse approximation for the leading effects of other more realistic profit
functions of limit orders. Both propositions 3 and 4 are illustrated in the graph. In particular, as transmission latency drops, the empirically examined order book depth does not necessarily monotonically improve or worsen. Depending on the parametrization of $\pi(\cdot)$, the resulting shapes of the figures vary qualitatively.

### 4.4 Policy discussion

The above model-predicted order book depth dynamics captures some stylized empirical facts, such as clustered order submission followed by immediate massive cancellation. Such behavior is similar in appearance to the ill-purposed trading strategy known as “quote stuffing”. For example, Egginton, van Ness, and van Ness (2012) refer to quote stuffing as “a practice in which a large number of orders to buy or sell securities are placed and then canceled almost immediately” and similar definitions are found in the media.

The model lends a rather innocuous explanation to the so-called quote stuffing behavior. The overshoot and the following immediate cancellation may simply be an equilibrium outcome due to queuing uncertainty, a market imperfection of low latency trading. This view echoes with Baruch and Glosten (2013)’s model prediction of “flickering quotes” in a zero-tick order book but with a different mechanism. In particular, the current paper focuses on the book depth dynamics, or “flickering depth”.

Along this line, this paper further argues that the recent regulation on limiting order-to-trade ratio and on minimum quote life is worth debating. For example, Oslo Børs has imposed a bound of 70 on the ratio of monthly orders to executed orders for each member, with an additional charge for exceeding this amount. Borsa Italiana has also implemented similar a order-to-trader policy in 2012. The London Stock Exchange imposed and has been adjusting a “high usage surcharge” on orders and quotes since 2010.

Such regulations limit market makers’ revision opportunities. In response, the market makers can be expected to reduce their order sizes in the short-run, leading to a thinner, less liquid order book. The MiFID II proposal rightfully expresses the concern that “to set out the maximum ratio of unexecuted orders to transactions” should take “into account the liquidity of the financial instrument” (Article 51.7). The U.K. Foresight Projects also feature articles by Brogaard (2011), Friederich and Payne (2012), and Farmer and Skouras (2012), concerning similar negative effects on market quality. In particular, preliminary empirical evidence from Friederich and Payne (2012) finds that book depth in Italian stocks fell after an order-to-trade ratio limit is imposed.

---

5 Endogenous market order and welfare

The queuing uncertainty escalates the level of competition among market makers and hence pauperizes them. However, the overshoot benefits the liquidity demand side as the fundamental investors can trade to meet their liquidity needs at lower costs. This section adopts the fully endogenized framework to evaluate the overall effect of queuing uncertainty on social welfare. The analysis clarifies 1) how welfare depends on the liquidity in a limit order market and 2) under what conditions can the exchange improve welfare by randomizing the order queues.

Setup. The market settings are exactly the same as in Biais, Martimort, and Rochet (2000), except that queuing uncertainty is introduced and, for clarity, the focus is restricted to only one price level, $a$, the best ask price. The one-price focus is motivated by the fact that in reality, there are few market orders that are large enough to bite into the book. This way, the liquidity measure can be summarized in one scalar variable, the book depth. (With multiple prices, a complex “liquidity” measure is needed to account for variations in both price and depth.)

There are two types of agents: One investor with CARA utility and $n \geq 1$ risk-neutral market makers as described in section 3.1. The traded asset pays off $V$ units of the numéraire good upon consumption. The payoff $V$ is defined as $V = v_0 + S + \epsilon$, where $v_0$ is the unconditional expected payoff, $S$ is the investor’s private signal with $\mathbb{E}S = 0$, and $\epsilon$ is the unobservable innovation, assumed to be normally distributed with mean zero. Denote by $\rho (> 0)$ the product of the investor’s constant absolute risk-aversion coefficient and the variance of $\epsilon$. Timing of the game is as described in section 3.1. The investor observes the signal $S = s$ and also suffers from an endowment shock $E = e$, where both realizations $s$ and $e$ are her private information. Without loss of generality, let $\mathbb{E}E = 0$. The joint distribution of $E$ and $S$ is commonly known. The stochastic queue of limit orders, the innovation $\epsilon$, and the joint distribution of $E$ and $S$ are all independent.

\[\text{16} \]
To focus on the short-run order submission and to keep the model analytically tractable, unlike section 4, the game is only one-period as in Biais, Martimort, and Rochet (2000).
5.1 Equilibrium

Consider the investor’s optimization problem given a depth level \( y \) at the ask price \( a \). Conditional on her private information of \( S = s \) and \( E = e \), the investor optimizes her certainty equivalent:

\[
\max_{0 \leq x \leq y} m + (v_0 + \theta - a) x - \rho y x - \frac{\rho}{2} x^2
\]

where \( m := (s + v_0)(e + y) - \rho \cdot (e + y)^2/2 \) is the certainty equivalent for her endowment and \( \theta := s - \rho e \) is her endowment-risk-adjusted signal for the asset. With the lower and the upper bound of \( 0 \leq x \leq y \), the potentially cornered solution is

\[
(8) \quad x(\theta; y) = \min \left\{ y, \max \left\{ 0, \frac{1}{\rho} (v_0 + \theta - a) - y \right\} \right\}.
\]

It can be seen that the investor’s optimal order size truthfully reveals \( \theta \), an imperfect signal (due to the liquidity/endowment shock; see also Vayanos and Wang, 2012), up to the lower and the upper corners.

Consider next market makers’ profitability. Market makers do not know the investor’s (endowment-risk-adjusted) signal, \( \Theta \), but they know \( \Theta := S - \rho E \), hence also its c.d.f., denoted by \( F(\theta) \) (the joint density of \( S \) and \( E \) are common knowledge). Suppose the depth at this price \( a \) is \( y \geq 0 \). Then the expected profitability for these \( y \) units of the limit order is

\[
(9) \quad \pi(y) = \int_{a-v_0}^{\infty} (a - v(\theta)) x(\theta; y) dF(\theta) + \int_{a-v_0+\rho y}^{\infty} (a - v(\theta)) y dF(\theta)
\]

where \( x(\theta; y) \) is the investor’s market order size at this price level (equation 8), and \( v(\theta) := \mathbb{E}[V | \Theta = \theta] \) is the conditional expectation of the asset payoff. Note that the integration begins from \( a - v_0 \) as \( x(\theta; y) \) is non-zero only if \( \theta > a - v_0 \), i.e. only if the signal is strong enough. Following Biais, Martimort, and Rochet (2000), two regularity conditions are needed. First, the joint distribution of \( S \) and \( E \) satisfies that \( v(\theta) \) is weakly increasing in \( \theta \), i.e. \( \dot{v}(\theta) \geq 0 \). Second, the support of \( \Theta \) is continuous.

The marginal expected profit is

\[
(10) \quad \pi(y) = \int_{a-v_0+\rho y}^{\infty} (a - v(\theta)) f(\theta) d\theta = \mathbb{P}(\Theta > a - v_0 + \rho y)(a - \mathbb{E}[v(\Theta) | \Theta > a - v_0 + \rho y]).
\]

The second equality gives the intuitive interpretation of the expected profit of the marginal order: Conditional on its execution (with probability \( \mathbb{P}(\Theta > a - v_0 + \rho y) \)), the sales revenue is the ask price \( a \) and the adverse-selection cost is \( \mathbb{E}[v(\Theta) | \Theta > a - v_0 + \rho y] \). The following proposition shows that “top-of-queue
advantage” holds (and, in fact, is further strengthened).

**Proposition 5 (Top-of-queue advantage).** There is top-of-queue advantage. In particular, both assumptions 1 and 2 hold for a large enough. Further, \( \pi(y) \) is quasi-convex in \( y \geq 0 \).

(A version of the top-of-queue advantage deep in the book is given later in section 6.) As soon as \( \dot{\pi}(0) > 0 \), which is always true when the best ask price \( a \) is large, the analysis in the baseline model (section 3) applies, and the equilibrium order submission strategy is characterized by the first-order condition (2) and the existence is given by lemma 1.

### 5.2 Liquidity provision and welfare

Define welfare as the sum of the (ex ante) expected gains from trade of all participants, the investor and \( n \) market makers, measured by their respective ex ante certainty equivalents. Suppose the book depth is \( y \). The investor’s certainty equivalent is 
\[
\text{ce}(y) = \mathbb{E}[M + (v_0 + \Theta - a) x(\Theta; y) - \rho x(\Theta; y)^2 / 2],
\]
where \( M := E \cdot (S + v_0) - \rho E^2 / 2 \) is the certainty equivalent of the investor’s endowment, \( \Theta := S - \rho E \) is the endowment-risk-adjusted signal, and \( x(\Theta; y) \) is the investor’s optimal market order size (which might be capped; see equation 14). The expected gains from trade is
\[
\text{ce}(y) - \mathbb{E}M = \mathbb{E} \left[ (\Theta - [\theta]) x(\Theta; y) - \frac{\rho}{2} x(\Theta; y)^2 \right],
\]
where \( [\theta] := a - v_0 \) is the floor which \( \Theta \) must exceed so that the investor will trade.

The market makers are risk-neutral and their aggregate expected certainty equivalent is just the sum of their expected profits.

**Lemma 5 (Market makers’ aggregate profit).** The aggregate profit of all market makers is queue-realization irrelevant. Only the aggregate book depth matters: \( \forall k \in \mathcal{K}, \sum_i \pi(Q_i^k + q_i) = \pi(\sum_i q_i) \).

With lemma 5, the market makers’ aggregate expected gains from trade, fixing a book depth \( y \), can be computed as (c.f. equation 9)
\[
\pi(y) = \mathbb{E} \left[ ([\theta] + v_0 - v(\Theta)) x(\Theta; y) \right].
\]

Therefore, the social welfare, \( (\text{ce}(y) - \mathbb{E}M) + \pi(y) \), is a function of the book depth, i.e. the liquidity level in
this market:

\[ w(y) = \mathbb{E} \left[ (v_0 + \Theta) x(\Theta; y) - \frac{\rho}{2} x(\Theta; y)^2 - v(\Theta) x(\Theta; y) \right]. \tag{11} \]

Note that the transfer, \([\theta] x(\Theta; y)\), from the market makers (adverse selection cost) to the investor (information rent) offsets in the aggregation. The remaining terms have intuitive interpretations. The first term is the investor’s expected valuation of the position she buys from the market makers. The second term corrects her acquired risk. The last term is the value of the position seen from market makers’ perspective (recall that \(v(\theta) := \mathbb{E} [V | \Theta = \theta] \)).

**Lemma 6 (Shape of welfare as a function of liquidity).** For all \(y \geq 0\), \(w(y) > \pi(y)\), \(\dot{w}(y) > \dot{\pi}(y)\), and \(\ddot{w}(y) < \ddot{\pi}(y)\). Further, \(w(y)\) is concavely monotone increasing on \(y \in [0, y^\pi]\).

Lemma 6 only says that welfare is initially increasing. There might be, however, times when too much liquidity hurts social welfare. To see this, write the derivative of \(w(y)\) as

\[ \dot{w}(y) = \mathbb{P}(\Theta > [\theta] + \rho y) (\mathbb{E} [v_0 + \Theta - \rho y | \Theta > [\theta] + \rho y] - \mathbb{E} [v(\Theta) | \Theta > [\theta] + \rho y]). \]

The difference between the two conditional expectations is the wedge between the investor’s and market makers’ valuation of the marginal unit of the asset. The difference is not necessarily positive as, from the social planner’s perspective, the investor might buy “too aggressively”. This is because though the investor enjoys her information rent, the social planner does not, as such rent offsets with the market makers’ adverse-selection cost. Consider the extreme case where the investor is very likely to draw a positive signal but no endowment shock: In this case, large liquidity provision in the market is socially suboptimal because it “indulges” the risk-averse agent to buy risky asset from the risk-neutral market makers. Put alternatively, the market makers’ competition, due to queuing uncertainty, might generate negative externality as the liquidity overshoot indulges inefficient reallocation of risky assets. Figure 5 shows an initially increasing welfare function eventually decreases if there is too much liquidity. The parametrization for the numerical procedure is described in section 5.4.

**Proposition 6 (Maximum social welfare).** Social welfare maximizes at \(y^*\), where \(y^*\) is strictly larger than \(y^\pi\) (market makers’ break-even depth) and is possibly infinite.
Figure 5: Welfare as a function of liquidity supply. This figure qualitatively illustrates the shape of welfare as a function of aggregate liquidity supply (i.e. aggregate book depth). The numerical procedure is described in section 5.4. The vertical axis shows welfare, market makers’ expected profit, and the investor’s certainty equivalent. The horizontal axis shows the book depth, $y$. The upper curve (blue) is welfare, $w(y)$. The lower curve (red) is market makers’ aggregate expected profit, $\pi(y)$. The difference, indicated by the (green) vertical line, is the investor’s certainty equivalent, $ce(y)$.

Compare the cost-benefit analyses of a social planner and of a market maker:

$$
\hat{w}(y) - \hat{\pi}(y) = \int_{[\theta] + [\theta]}^\infty \left[ (v_0 + \theta - \rho y) - a \right] f(\theta) d\theta
$$

$$
= P(\Theta > [\theta] + \rho y) \left( E[v_0 + \Theta - \rho y | \Theta > [\theta] + \rho y] - a \right) > 0.
$$

The difference is exactly the investor’s expected benefit of buying a marginal unit of the asset. (If she could buy the marginal unit, she gains the conditional expectation term and she pays the ask price $a$.) That is, compared to a representative market maker who chooses the book depth to maximize the aggregate expected profit, a social planner in addition takes into account the investor’s additional gain. The socially optimal depth, therefore, always exceeds the market making optimal depth. The result is analogous to the scenario of a monopolist producer who, facing an exogenous demand function, chooses the production quantity to maximize his profit. There is always some consumer surplus cannot be realized because of the monopolist under-supplies in order to sell at a high price.

Recall that with queuing uncertainty, the order book deepens (liquidity overshoot) and hence there is
room for welfare improvement by potentially adding queuing uncertainty to the market. This leads to the following subsection, which introduces queue randomization and analyzes how welfare, i.e. aggregate expected gains from trade, is affected.

5.3 Market design

In this subsection, consider \( n = 2 \) (representative) market makers, a fast one and a slow one. The two-agent assumption is an abstraction to reflect speed heterogeneity in the real-world trading environment, where some participants, having invested heavily in technology, have relatively low latency accessing the market compared to others. The two-agent restriction is imposed for the purpose of tractability. The numerical procedure (discussed in section 5.1) suggests that the results in this subsection hold generally.

Label the two market makers by 1 and 2, and there are two possible queues of the limit orders: \( k \in \mathcal{K} = \{[1, 2]^T, [2, 1]^T\} \). Denote by \( \alpha := \mathbb{P}(K = [1, 2]^T) \) the probability of market maker 1’s order queuing in the front of market maker 2’s. Without loss of generality, let market maker 1 be the (weakly) faster one: \( \alpha \in (1/2, 1) \). In equilibrium, the limit order sizes, \( q_1 \) and \( q_2 \), should satisfy the following first-order condition system (c.f. equation 2):

\[
\begin{align*}
\dot{q}_1 + (1 - \alpha)\dot{q}_2 + q_1 &= 0, \quad \text{for market maker 1;} \\
(1 - \alpha)\dot{q}_2 + \alpha \dot{q}_1 + q_2 &= 0, \quad \text{for market maker 2.}
\end{align*}
\]

(12)

The following technical condition is needed for comparative statics:\(^{17}\)

\[
\dot{\pi}(2y^\alpha) \leq 0.
\]

(13)

Then a connection between the aggregate book depth and speed heterogeneity can be established:

**Proposition 7 (Speed heterogeneity and liquidity overshoot).** The liquidity overshoot is more pronounced when the market makers have heterogeneous speed. Mathematically, if denote the equilibrium aggregate book depth by \( y \), then \( y(\alpha) \) has a local maximum at \( \alpha = 1/2 \).

---

\(^{17}\) Condition (13) ensures that \( \pi(\cdot) \) is concave on a support large enough. Effectively, it ensures the equilibrium is a “stable” one. The importance of a stable equilibrium for comparative statics has been emphasized in Samuelson (1941, 1942).
Corollary 1. If \( \hat{\pi}(y) \) is “regular”, the equilibrium aggregate book depth \( y \) is monotone increasing (decreasing) on \( \alpha \in (0, 1/2) \) (on \( \alpha \in (1/2, 1) \), respectively) and has its unique maximum at \( \alpha = 1/2 \).

Remark. A “regular” \( \hat{\pi}(\cdot) \) is important for the purpose of comparative statics. This is because the market makers’ problem, as alluded in proposition 2, resembles firms’ competition game and inherits the difficulty of signing comparative statics from such oligopolistic competition games where agents’ decisions are strategic substitutes of each other. A convenient (sufficient) condition is that \( \hat{\pi}(\cdot) \) is concavely decreasing (c.f. a concavely decreasing demand function as in Kreps and Scheinkman, 1983) on the relevant support, as shown in the proof. When concavity does not hold, the proof shows that if the higher-order effects are small, the proposition is still valid.

“Speed” in the context is reflected in the probability of queuing in the front of the queue, and it is reflected by parameter \( \alpha \) for market maker 1 and by \( 1 - \alpha \) for market maker 2. When \( \alpha = 1 - \alpha = 1/2 \), there is no speed heterogeneity. As \( \alpha \) increases towards 1 (or decreases towards 0), the speed heterogeneity increases. By symmetry, the proposition implies that the overshoot reaches its maximum at \( \alpha = 1/2 \) and then monotonically decreases. Conceptually, therefore, less speed heterogeneity is equivalent to more queuing uncertainty, hence also more fierce competition among market makers. When there is no speed heterogeneity (most queuing uncertainty), the competition is most fierce and the most pronounced liquidity overshoot follows. When \( \alpha \to 1 \), market maker 1 is almost surely to be the first in the queue and will submit a limit order of size \( y^= \) to maximize his expected profit. The competition vanishes in this extreme case and there is no overshoot (setting \( \alpha \) to 1 in the first-order condition system 12).

Now consider adjusting the queuing uncertainty in the market by introducing a queue randomizer, \( R \), defined as a square random matrix

\[
R = \begin{bmatrix}
B & 1 - B \\
1 - B & B
\end{bmatrix},
\]

where \( B \) is a Bernoulli random variable with success probability \( \beta \in [0, 1] \). For example, the exchange can apply the queue randomizer to every realized queue in a fixed time interval. Specifically, \( R \) applies to a realized queue, \( k \), by pre-multiplying it: The randomized queue becomes \( K_R = Rk \), which is either \( K_R = k \) (with probability \( \beta \)) or \( K_R \in \mathcal{K} \setminus \{k\} \) (with probability \( 1 - \beta \)). Therefore, when the exchange adopts a queue
randomizer $R$, the effective queue distribution can be characterized by

$$\alpha_R(\beta) := \mathbb{P}(K_R = [1, 2]^T) = \alpha \beta + (1 - \alpha)(1 - \beta) \in [1 - \alpha, \alpha],$$

thanks to the simplified assumption that $n = 2$. When there are $n \geq 3$ market makers, the queue distribution is more complex to characterize. Appendix A imposes additional structure to characterize the queue distribution. Section 5.4 adopts the micro-foundation in appendix A to numerically illustrate the policy implication.

**Corollary 2 (Optimal queue randomizer).** There exists a randomizer $R$ such that welfare can be (weakly) improved when $R$ is applied.

Effectively, the queue randomizer adjusts the queuing uncertainty (equivalently, speed heterogeneity) and also the degree of competition between the market makers. For example, consider the case where there is very severe speed heterogeneity, i.e. little queuing uncertainty. Without queue randomization, $y$ is close to $y^*$ (lack of competition). By proposition 6, welfare can be improved by increasing the equilibrium book depth in this case. One way to do so is to add more queuing uncertainty by, for example, setting $\beta \leq 1/2$ so that the originally slow market makers are more likely to be re-queued to the front. This randomization curtails the “market power” of the fast market makers and strengthens the competition, which results in larger overshoot and a deeper book in equilibrium, benefiting the investors and improving welfare.

It should be noted, however, such queue randomization must be implemented with care for at least three reasons. First, the effectiveness of queue randomization is bounded by the initial speed heterogeneity among market makers. If there is little speed heterogeneity ($\alpha \to 1/2$), the queue randomization is ineffective ($\alpha_R \to 1/2$).

Second, improperly adjusting the queuing uncertainty can destroy social welfare. For example, suppose the unrandomized equilibrium aggregate book depth $y(\alpha)$ (with $\alpha \geq 1/2$) already exceeds the socially optimal depth $y^*$: $y^* < y(\alpha)$. In this case, improper queue randomization will reduce welfare because by adding more queuing uncertainty, the overshoot exacerbates, $y(\alpha_R(\beta)) \geq y(\alpha)$, and the increased liquidity supply is not appreciated by the social planner (see lemma 6 and its discussion).

Finally, by adding queuing uncertainty, the exchange essentially deprives the market power from fast market makers (who invest heavily in technology to achieve such speed advantage). The reduced profitability
might, over time, constrain the participation of market makers. Such concerns, though beyond the scope of the current model, are legitimate and should be accounted for in choosing optimal queue randomization schemes.

5.4 Numerical example

This subsection adds normality to the model, following the example of Glosten (1994) and also example 1 of Back and Baruch (2013). Let the investor’s private signal and endowment shock be independent and both normally distributed with zero means. The endowment-risk adjusted signal is \( \Theta = S - \rho E \), which is then also normally distributed with zero mean and variance \( \text{var}\Theta := \text{var}S + \rho^2\text{var}E \). The market maker’s inference, given a signal \( \theta \), about the asset value is \( v(\theta) = v_0 + \gamma \theta \), where the price impact factor is \( 0 < \gamma = \text{var}S / \text{var}\Theta < 1 \). In the numerical illustration below, the parameters are \( a = v_0 + 1, \text{var}S = 0.2, \text{var}E = 0.1, \) and \( \rho = 0.9 \).

The investor’s certainty equivalent, the market makers’ aggregate expected profit, and welfare can then be evaluated using the expressions derived in section 5.2. Figure 5 illustrates the shape of welfare as a function of the aggregate book depth, contrasting with the market maker’s aggregate expected profit. It can be seen that welfare is not monotone increasing in liquidity supply in this example. Further, when market makers’ expected profit maximizes, welfare can still be improved by increasing the liquidity supply from \( y^* \) (\( \approx 0.2495 \)) to \( y^+ \) (\( \approx 0.5216 \)).

When there is queuing uncertainty (and multiple market makers), the equilibrium book depth exceeds \( y^* \) and hence, there is room for welfare improvement by adding queuing uncertainty. However, when there is too much queuing uncertainty, the liquidity overshoot will be so excessive that welfare reduces. This would be the case when the equilibrium book depth exceeds \( y^+ \).

To visualize the effect directly, the following parametrization is prepared for figure 6. Let each market maker be endowed with a type \( \mu_i \), such that the expected waiting time (latency) for market maker \( i \)'s order to be processed by the exchange server is \( E L_i = \mu_i \), where \( L_i \) is exponentially distributed. The parameter \( \mu_i \) is interpreted as the (expected) “latency” of market maker \( i \); a faster market maker has a smaller \( \mu_i \). Assume \( L_i \) is independent of \( L_j \) if \( i \neq j \). Appendix A gives the details and provides closed-form solutions to the queue distribution \( \mathbb{P}(K) \). Finally, assume that the \( n \) market makers all have heterogeneous speed: \( \mu_i - \mu_{i+1} = \Delta \mu > 0 \) for all \( i \in \{1, ..., n - 1\} \); that is, the speed difference is equally spaced. Note that, fixing \( n \), increases in \( \Delta \mu \)
Figure 6: Queuing uncertainty and welfare. This figure illustrates how welfare and the equilibrium book depth respond to the level of queuing uncertainty in the market. The red curve is the equilibrium book depth ($y$, left axis) and the blue curve is welfare ($w$, right axis). The horizontal axis represents the level of queuing uncertainty, defined as the speed difference between market makers ($\Delta \mu$). The parameters are as described in section 5.4.

implies aggravated speed heterogeneity in market makers’ speed.\textsuperscript{18}

Figure 6 plots the book depth (liquidity supply) and welfare in response of the level of queuing uncertainty in the market. The number of market makers is $n = 5$. The fastest market maker’s speed is normalized to $\mu_1 = 1$ and $\Delta \mu$ ranges from 0 to 1.8. It can be seen that as queuing uncertainty increases, though the equilibrium book depth monotonically increases, welfare is first improved and then destroyed. The shape of the welfare curve is consistent with figure 5.

6 Deep in the book

This section studies the consequence of queuing uncertainty deep in the book. It shows the robustness of liquidity overshoot result (proposition 1), the driver of the order book depth dynamics in section 4 and the welfare analysis in section 5.

\textsuperscript{18} The parametrization adopted in this subsection for speed heterogeneity is one of the many possibilities. Other alternatives have been experimented and are shown to generate qualitatively similar patterns as shown in figures 5 and 6.
Model setup with discrete prices. The model extends the setup in section 5 so that multiple prices are possible. Unlike Biais, Martimort, and Rochet (2000, 2013), in order to make sense of queuing uncertainty, prices in the order book must be discrete. Let \( \mathcal{P} \) denote the collection of all possible price levels in the limit order book: \( \mathcal{P} = \{ p_{\min}, \ldots, p_{\max} \} \), where \(-\infty < p_{\min} < v_0 < p_{\max} < \infty \). Describe the book depth by \( y(p) : \mathcal{P} \mapsto \mathbb{R} \), which gives the (signed) depth at price level \( p \). As before, because the analysis is symmetric, this paper only considers the ask side. Denote the \( j \)-th ask price by \( a_j \) (\( j \in \{1, 2, \ldots\} \)) which is the \( j \)-th smallest price in \( \{ p \mid p > v_0, p \in \mathcal{P}, y(p) > 0 \} \).\(^{19} \) Let the depth at the \( j \)-th best price be denoted by \( y_j \) and define \( a_0 = v_0 \) with \( y_0 = 0 \). (It will be justified that the ask side indeed corresponds to the prices \( \{ p \mid p > v_0, p \in \mathcal{P} \} \) in equilibrium.) Finally, denote the cumulative depth up to and including (excluding) the \( j \)-th level by \( y_{\leq j} := \sum_{i \leq j} y_i \) (and \( y_{< j} := \sum_{i < j} y_i \) similarly).

6.1 The investor’s order size

Analogous to the analysis in section 5.1, consider the investor’s optimization problem given a book described by \( y(p) \). Suppose the investor wants to buy the asset offered at the \( j \)-th ask price, with price \( a_j \) and depth \( y_j \) (> 0). Note that, the optimization is iterative: Due to price priority, the investor can only buy the offers at the \( j \)-th ask price if she has already bought everything offered below the \( j \)-th price. That is, she must have bought \( y_{< j} \) units of the asset, accumulating her endowment to \( e + y_{< j} \) units, before she could choose her optimal demand at price \( a_j \). Then the investor optimizes:

\[
\max_{0 \leq x_j \leq y_j} \mathbb{E} \left[ -\exp \left\{ -\left[ \left( E + y_{< j} \right) V + \left( V - a_j \right) x_j \right] \right\} \mid E = e, S = s \right]
\]

which is equivalent to maximizing her certainty equivalent:

\[
\max_{0 \leq x_j \leq y_j} \ m_j + \left( v_0 + \theta - a_j \right) x_j - \rho y_{< j} x_j - \frac{\rho}{2} x_j^2
\]

where \( m_j := (s + v_0)(e + y_{< j}) - \rho (e + y_{< j})^2 / 2 \) is the certainty equivalent for her holding (after buying everything offered below \( a_j \)) and \( \theta := s - \rho e \) is her endowment-risk-adjusted signal for the asset. (Note that the transfer, \( \sum_{i < j} y_i a_i \), that the investor has already paid does not come into the optimization.) The potentially cornered

---

\(^{19}\) Perhaps surprisingly, under queuing uncertainty, it is not necessary that the equilibrium ask prices are consecutive; i.e. it is possible that \( a_{j+1} - a_j > p \). See corollary 3 for details.
The investor’s optimal market order size. The investor’s cumulative order size, \( x_{\text{cor}} := \sum_j x_j \), is plotted against her endowment-risk-adjusted signal, \( \theta \), given a limit order book with depths \( y_j \) at the \( j \)-th best ask price. The curve is piece-wise increasing in \( \theta \). The investor bites into price level \( a_j \) only if \( \theta \) exceeds a floor threshold \( [\theta_j] \) and depletes that level only if \( \theta \) exceeds a ceiling \( [\theta_j] \) as given in lemma 7. The hurdle size is \( a_{j+1} - a_j \).

solution is

\[
   x_j(\theta; y_j) = \min \left\{ y_j, \max \left\{ 0, \frac{1}{\rho} \left( v_0 + \theta - a_j \right) - y_{<j} \right\} \right\},
\]

Clearly, equation (8) is a special case of the more general solution above. The corners give rise to the following “hurdles” that do not exist if the price is continuous:

**Lemma 7 (Hurdle in signal).** After the book depth at the \( j \)-th best price depletes, the investor further bites to the \( (j+1) \)-th price level if and only if her (endowment-risk-adjusted) private signal exceeds a strictly positive hurdle.

Mathematically, for each ask price \( a_j \), there exist a floor \( [\theta_j] = (a_j - v_0) + \rho y_{<j} \) and a ceiling \( [\theta_j] = (a_j - v_0) + \rho y_{<j} \) such that \( x_j(\theta) \) is off-corner if and only if \( [\theta_j] < \theta \leq [\theta_j] \). The hurdle between two consecutive price levels is \( [\theta_{j+1}] - [\theta_j] = a_{j+1} - a_j > 0 \).

Figure 7 illustrates the investor’s optimal cumulative order size as a function of her endowment-risk-adjusted signal \( \theta \) and the book \( \{y_j\} \). This cumulative order size is essentially her demand for the asset (given the book depth), which is piece-wise linearly increasing with slope \( 1/\rho \) (the risk-adjustment).
6.2 Depths in the order book

Consider the $j$-th ask price level at $a_j = p > v_0$, with the previous cumulative depth $y_{<j}$. Suppose the depth at this price $a_j$ is $y \geq 0$ and then the expected profitability for these $y$ units of the limit order is

$$\pi_j(y) = \int_{[\theta_j]}^{[\theta_j] + py} (a_j - v(\theta) - v(\theta)) x_j(\theta) f(\theta) d\theta + \int_{[\theta_j] + py}^{\infty} (a_j - v(\theta)) y f(\theta) d\theta$$

where the signal floor $[\theta_j]$ is defined in lemma 7, $x_j(\theta)$ is the investor’s market order size at this price level (equation 14), $v(\theta) := \mathbb{E}[V|\Theta = \theta]$ is the conditional expectation of the asset payoff, and $f(\theta)$ is the density of $\Theta$ (the density is assumed to exist, following Biais, Martimort, and Rochet, 2000). The marginal expected profit is

$$\dot{\pi}_j(y) = \int_{[\theta_j] + py}^{\infty} (a_j - v(\theta)) f(\theta) d\theta.$$  

The profit and marginal profit expressions (15) and (16) generalize the one-price level versions in equations (9) and (10), respectively. The following is a generalized version of proposition 5 seen in section 5.1.

**Proposition (Top-of-queue advantage).** There is top-of-queue advantage in the limit orders at each price level. In particular, both assumptions 1 and 2 hold. Further, $\dot{\pi}_j(y)$ is quasi-convex in $y \geq 0$.

Now turn to the market makers. Each market maker submits a set of limit orders, one order for each price level, to maximize his expected profit. Because the cumulative depth at any price level affects the profitability of orders deep in the book, the optimization is dynamic (along the price dimension) in nature. Consider market maker $i$’s order size at price level $j$, given all other market makers’ order sizes. Denote by $u_{i,j}$ his continuation value for price level $j$ and the Bellman equation can be written as

$$u_{i,j} = \max_{q_{i,j}} \mathbb{E}\left[\pi_j(Q_{i,j}^q + q_{i,j}) - \pi_j(Q_{i,j}^q)\right] + u_{i,j + 1},$$

where for notation simplicity, the arguments of $u_{i,j+1}$ are omitted but it should be understood that it depends on market maker $i$’s order size $q_{i,j}$ at price level $j$. With some manipulation, it can be shown that the first-order condition can be written as (see the proof of proposition 8 in appendix)

$$\mathbb{E}\pi_j(Q_{i,j}^q + q_{i,j}) - \mathbb{E}\pi_{j+1}(Q_{i,j+1}^q) = 0.$$
Compared with the first-order condition in the baseline model (equation 2), there appears a new term, $-\mathbb{E} \pi_{i,j+1}(Q^i_{j,j+1})$, in the equation. This new term measures the expected profit of the first marginal unit of the order on price level $j + 1$. It is negative because if market maker $i$ increases his order size at level $j$, the profitability at level $j + 1$ reduces (lower execution probability and higher adverse-selection cost). The interpretation of this first-order condition is, instead of myopically choosing the order size to break even the expected profit at the current price level, market maker $i$ also accounts for his expected profitability on the next price level. This effect, however, is not large enough to curb the fierce competition among market makers and the aggregate liquidity at each price level still overshoots, as proved in the following proposition.

**Proposition 8 (Overshoot deep in the book).** Suppose the investors’ endowment-risk-adjusted signal $\Theta$ exhibits (weakly) increasing hazard rate. Then, in equilibrium, there is liquidity overshoot at each price level. Mathematically, $\pi_j(\sum_i q_{i,j}) \leq 0$ for all $j$ and the equality holds if and only if there is one market maker who is almost surely the first in queue at price level $j$.

**Remark.** The technical condition that $\Theta$ exhibits (weakly) increasing hazard rate, i.e. $f(\theta)/(1 - F(\theta))$ is weakly increasing in $\theta$, is a sufficient (but not necessary) condition for the overshoot result. As discussed in Biais, Martimort, and Rochet (2000), who also assume such a condition (equations 16 and 17), this is not a restrictive assumption. Some commonly used distributions (to name a few, uniform, normal, and exponential) all satisfy increasing hazard rate. The hazard rate essentially reflects the severity of market makers’ adverse-selection problem. Therefore, the increasing hazard rate assumption can also be understood, intuitively, as a condition to ensure the regularity of the problem; see Baruch and Glosten (2013) for more details on the importance of sufficient adverse-selection in the limit order market.

The overshoot results suggest that instead of break-even, the last unit limit order at each price level loses in expectation, at least in very short run. This gives an equilibrium prediction about “holes” in the limit order book (in short-run):

**Corollary 3 (Holes in book).** For sufficiently small tick size, the equilibrium limit order book has, possibly interior, “holes”, or empty price levels on which no market makers are willing to post any orders.

To understand the intuition of this result, it is helpful to look at again the expression of the expected marginal
profit (equation 18). When the price level goes from \( j \) to \( j + 1 \), the (expected) marginal revenue jumps from \( a_j \) to \( a_{j+1} \) and, yet, the expected marginal cost, which only depends on the cumulative book depth, does not change. Liquidity overshoot implies that the expected marginal cost of the last limit order at price level \( j \) exceeds \( a_j \). Hence, if the next price level \( a_j + \rho \) (where \( \rho \) is the tick size) is not large enough, the marginal profit at price \( a_j + \rho \) will never be positive and no market maker will place order there.

Empirical evidence of holes in limit order book has been found by, for example, Biais, Hillion, and Spatt (1995). Queuing uncertainty gives an equilibrium explanation. In particular, equilibrium holes are not supported by the equilibrium of Glosten (1994), whose break-even condition asserts that the last unit of limit orders at each price level has expected revenue equating expected cost. This is because, as the price level moves to the next tick, the expected revenue jumps up but the expected cost does not change, implying strictly positive expected marginal profit for the first unit of limit orders at the new price. Market makers will compete for such profit by submitting orders to the new level, and there will be no holes.

It should be emphasized that the predicted holes in the book are *transitory* in nature. If the market makers are given sufficient time to revise/cancel their orders, the long-run equilibrium reverts to the break-even level described by Glosten (1994) and the holes will disappear; see section 4 for the book depth dynamics.

A second corollary follows proposition 8. It compares the cumulative liquidity in the book under queuing uncertainty and under the break-even condition of Glosten (1994).

**Corollary 4 (Cumulative liquidity overshoot).** Denote by \( y^*_{\leq j} \) and \( y^\pi_{\leq j} \), respectively, the equilibrium cumulative depths until (inclusive) the \( j \)-th price level under queuing uncertainty and under the break-even condition (Glosten, 1994). If there are no holes in the book under queuing uncertainty, then \( y^*_{\leq j} \geq y^\pi_{\leq j} \) for all price levels.

This corollary is not a restatement of proposition 8 because the proposition focuses on the depth at each price level while the corollary addresses the cumulative depth. In particular, the overshoot at price level \( j \) affects the break-even quantity at price level \( j + 1 \).

### 6.3 Alternative setups

So far this section has been building on the fully-endogenized framework of Biais, Martimort, and Rochet (2000). This subsection gives two other examples, following Back and Baruch (2013), to show that the
top-of-queue advantage (assumptions 1 and 2) holds generically.

**Example 1** (Informed or liquidity dichotomy). This example adopts the classical dichotomy that treats the incoming market order to be driven by either private information or liquidity shock. Such dichotomy is seen in, for example, Glosten and Milgrom (1985), Kyle (1985), Easley and O’Hara (1992), and many others.

Suppose that with probability \( \psi \) the arriving investor observes the true value of the asset, \( v \), while with probability \( 1 - \psi \) the arriving investor is purely liquidity driven. If the investor is information driven, she buys all units of the asset on offer at prices \( p \) \( v \). If the investor is liquidity driven, she submits a market order of size \( X \), which has density \( f(x) \) defined on \( \mathbb{R} \). Then market makers’ expected profit of a limit order of size \( y \) at price level \( a_j \) is

\[
\pi_j(y) = \psi \cdot (a_j - \mathbb{E}[V | V \geq a_j]) y + (1 - \psi) (a_j - v_0) \left[ \int_{y_{<j}}^{y_{<j} + y} (x - y_{<j}) f(x)dx + \int_{y_{<j} + y}^{\infty} y f(x)dx \right]
\]

where, following the notations developed in this section, \( y_{<j} \) is the cumulative depth excluding the \( j \)-th price level. It is easy to evaluate that \( \pi_j(y) = -(1 - \phi)(a_j - v_0) f(y_{<j} + y) < 0 \); that is, \( \pi_j(y) \) is strictly concave at any price level \( a_j > v_0 \). Therefore, assumption 1 holds under this setup, with the quasi-concavity strengthened to concavity, and hence follows assumption 2.

**Example 2** (Inelastic demand). In this example, the incoming market order size follows an exogenous distribution that is independent of the order book depth. Such models are seen in, for example, Sandås (2001), Foucault and Menkveld (2008), and Kervel (2013).

Denote the market order size by \( X \), a random variable with density function \( f(x) \). Denote the price impact by \( \nu(x) := \mathbb{E}[V | X = x] \) and assume \( \nu(x) \geq 0 \). Then given the cumulative depth \( y_{<j} \), market makers’ expected profit of a limit order of size \( y \) at price \( a_j \) is

\[
\pi_j(y) = \int_{y_{<j}}^{y_{<j} + y} (a_j - \nu(x))(x - y_{<j}) f(x)dx + \int_{y_{<j} + y}^{\infty} (a_j - \nu(x)) y f(x)dx.
\]

The expression is akin to equation (15), the fully-endogenized version. Indeed, following the proof of proposition 5, it can be shown that \( \pi_j(y) \) is quasi-concave and, further, \( \pi_j(y) > 0 \) implies \( \pi_j(y) < 0 \) as in proposition 5, satisfying assumptions 1 and 2.

The other results derived in section 6.2 (proposition 8 and its two corollaries) are still valid under the setups of these two examples. The derivation and proofs are omitted for brevity.
7 Conclusion

In a high-speed trading environment, traders cannot perfectly condition their decisions on real-time market status. This is because between a market status update and a trader’s registering of the update, there is non-zero delay during which the market status is likely to evolve. This paper acknowledges such market imperfection by introducing queuing uncertainty to standard limit order market model. The agents strategically play a simultaneous game that results in liquidity overshoot in equilibrium because the queuing uncertainty intensifies the competition and weakens the strategic substitution effects of others’ decisions.

The model identifies transitory overshoot in liquidity provision as an equilibrium phenomenon in modern financial markets. Such liquidity overshoot results from fierce competition among liquidity suppliers (market makers) who are unable to perfectly condition their strategy on the random queue position of their limit orders. It is shown that such queuing uncertainty weakens the strategic substitution effects in the competition game.

Further, the paper argues that flickering orders and the so-called “ghost liquidity” naturally follow the overshoot of liquidity as liquidity providers cancel and revise those orders that end up in the bottom of the queue. The transitory nature of the liquidity overshoot is associated with two different types of latencies in the limit order market: reaction latency and transmission latency. Although reduction in either type of the latencies amplifies the transitory liquidity overshoot (because faster speed accessing the market guarantees better order revision opportunity and hence intensifies competition), the implications on the stabilization process of the limit order book are different. While drops in reaction latency prolong the stabilization (increased amount of revisions in a given time interval), drops in transmission latency hasten the stabilization (reduced update time for a given amount of revisions). The model predictions concur with recent empirical literature on how latency drops affect order book depth.

The model also adds to the recent debate on trading speed and related regulations. After endogenizing liquidity demanders (fundamental investors), the model examines market quality through the expected gains from trade of all agents in the economy and shows that the exchange can improve market quality by cautiously randomizing the order queues in the limit order book, effectively adjusting the level of queuing uncertainty. Caveats on the implementation of queue randomization are discussed.

This paper invites future empirical works to test the model’s prediction about order book depth dynam-
ics and to verify the source of liquidity improvement due to the advancement in trading technology. The theoretical framework developed in this paper also leaves room for further exploration of how particular regulations (for example, make/take fees, minimum order resting time, and etc.) may affect traders’ order submission and market quality under queuing uncertainty.
Appendix

A Speed and queuing

This appendix models how the stochastic queue, $K$, is realized according to market makers’ speed. The probability masses of potential $n!$ queuing outcomes are fully characterized by $n$ speed parameters of the market makers. The model specification captures a key feature that queue positions are relative in nature: Increasing all market makers’ speed by a same factor should not affect the distribution of the queue. Hence, “speed” in this paper should always be understood in a relative sense.

Endow each market maker $i$ with a speed $\mu_i > 0$. In what follows, $\mu$ will be interpreted as the (expected) latency and, interchangeably, its inverse, $\mu^{-1}$, will be referred to as the speed. Collate these speed parameters in a vector $\mu = [\mu_1, \ldots, \mu_n]^T$. Let Nature draw for each market maker a latency, $L_i (> 0)$, from a commonly known distribution with c.d.f. $G(l_i; \mu_i)$ where $l_i$ is defined on $(0, \infty)$. The latency draws are assumed to be independent. Order the realized latencies as $l_{i(1)} \leq \cdots \leq l_{i(n)}$, where $i(k)$ is the index of the market maker who draws the $k$-th smallest latency. Each market maker $j$’s queue position then realizes to be $K_j = k$ if $i(k) = j$.

The latency (or, simply, delay) measures how soon the market maker can “effectuate” the order size she chose in time 1. Such effectuation delay, in a limit order market, can be the waiting time for the order to travel from a market maker’s computer to the exchange server.

Assumption 5 (Latency distribution). Latency $L_i$ is exponentially distributed on $(0, \infty)$ with mean $\mu_i$ ($> 0$). That is, $\mathbb{P}(L_i \leq l) = 1 - e^{-l/\mu_i}$ for $l > 0$. $L_i$ is independent of $L_j$ for all $j \neq i$.

It is acknowledged that this assumption is not without loss of generality. First, it rules out simultaneity almost surely ($\mathbb{P}(L_i = L_j) = 0$ if $i \neq j$). Second, all market makers have strictly positive probability to be in any queue position, i.e. $\mathbb{P}(K_i = k) > 0$ for all $i$ and $k$. This rules out, for example, the case where market maker 1 is never the first. In exchange, assumption 5 brings tractability and the intuitive property that only relative speed matters for the distribution of the queue.

Probability mass function for the queue vector $K$ can be easily solved. Consider a queue realization $k = \{k_1, \ldots, k_n\}$ determined by latencies $L_{i(1)} < \cdots < L_{i(n)}$. The probability of realizing such a queue is

$$\mathbb{P}_\mu (K = k) = \mathbb{P}_\mu (L_{i(1)} < \cdots < L_{i(n)}) = \int_0^\infty \cdots \int_0^\infty dG (l_{i(n)}; \mu_{i(n)}) \cdots dG (l_{i(1)}; \mu_{i(1)}),$$

which under assumption 5 has closed-form solution

$$\mathbb{P}_\mu (K = k) = \prod_{j=1}^n \frac{\mu^{-1}_{i(j)}}{\sum_{h=j}^{n} \mu^{-1}_{i(h)}}.$$

The subscript $\mu$ on $\mathbb{P}(\cdot)$ emphasizes the dependence of the probability measure $\mathbb{P}$ on market makers’ speed $\mu$. The following two lemmas will be useful in characterizing the queue distribution.

Lemma 8 (Speed relativity). Under assumption 5, only market makers’ relative speed matters for the queue distribution. Mathematically, $\mathbb{P}_\mu (K) = \mathbb{P}_{c\mu} (K)$ for any $c > 0$. 

44
Lemma 9 (Speed increase and queue distribution). Holding all other market makers’ speed constant, an increase in the speed of a market maker pushes him to the front of the queue and all others to the back (in a probability sense).

Mathematically, write $C_i$ as a diagonal matrix of size $n$ with the $i$-th diagonal term equal to $c > 1$ and all other diagonal terms equal to 1. Then, for any $i \neq j$, the distribution $\mathbb{P}_\mu(K_i)$ first-order stochastically dominates the distribution $\mathbb{P}_\mu(K_j)$ and the distribution $\mathbb{P}_\mu(K_j)$ is first-order stochastically dominated by the distribution $\mathbb{P}_{C_i \mu}(K_j)$.

Rather intuitively, lemma 8 says if all market makers’ speed are boosted by a same (strictly positive) factor, the resulting distribution of their queue does not change. That is, market makers’ latencies are ordinal. Lemma 9 qualifies the effect on queue distribution of an increase in market maker $i$’s speed. The vague idea that faster market makers are more likely to queue in the front is proved, under assumption 5, in terms of stochastic dominance.

B Notation summary

General notations:
- $i$: market maker index, $i \in \{1, \ldots, n\}$ (in section 4, $i \in [0, 1]$ under assumption 3).
- $k$: queue position index.
- $k$: queue realization, a vector of size $n$.
- $\mathcal{K}$: the collection of all possible queues realizations.
- $n$: the total number of market makers.
- $N$, the set of all market maker indices, $N = \{1, \ldots, n\}$.
- $\pi$ (also $\bar{\pi}, \hat{\pi}$, etc.): the aggregate expected profit (and its derivatives) of all limit orders.
- $q_i$: the limit order size of market maker $i$.
- $q$: the vector collecting all market makers limit order sizes, $q = [q_1, \ldots, q_n]'$ and $q_{-1} = q \setminus \{q_i\}$.
- $Q_i^\circ(k)$: the aggregate depth before (excluding) market maker $i$’s order, given the queue realization $k$.
- $y$: the book depth.
- $y^\circ$, the “break-even” book depth level at which $\pi(y^\circ) = 0$.

Additional notations in section 4:
- $\delta$: market makers’ reaction latency.
- $\eta$: transmission latency.
- $h$: the number of order revisions.
- $h^*$: the maximum number of order revisions.
- $\lambda(h; \delta)$: the inverse (discrete) hazard rate of $T$, defined as $(1 - G((h + 1)\delta))/G((h + 1)\delta) - G(h\delta))$.
- $q(h)$: the after-revision order size for the active market makers.
- $\bar{T}$: the time at which the displayed order book stabilizes.
- $T$: the (exponentially distributed) stochastic arriving time of the market order. It has c.d.f. $G(T)$ with support $[0, \infty)$.
- $g(h)$: the true book depth after the $h$-th revision.

Additional notations in section 5:
- $a$: the ask price.
- $\alpha$: the probability of market maker 1 being first in queue; only in section 5.3.
- $B$: a Bernoulli random variable with success probability $\beta$; only in section 5.3.
- $\epsilon$: the investor’s (ex ante) certainty equivalent.
- $E$: the endowment (ex ante) certainty equivalent.
- $\theta$: the density function of $\Theta$.
- $K_R$: the randomized queue, $K_R = RK$; only in section 5.3.
- $M$: the investor’s certainty equivalent of her endowment; $M := E \cdot (S + v_0) - \rho E^2/2$.
- $\rho$: the CARA investor’s risk-adjustment, the product of her absolute risk-aversion coefficient and the variance of $\epsilon$. 

45
- $R$: the queue randomizer, a 2-by-2 random matrix; only in section 5.3.
- $S$: the investor’s private signal about the asset value with zero mean.
- $\Theta$: the investor’s endowment-risk-adjusted marginal valuation of the asset; $\Theta := S - \rho E$.
- $V$: the asset value (payoff).
- $v_0$: the unconditional expected asset value, $v_0 = \mathbb{E}V$.
- $w$ (and $\dot{w}$, $\ddot{w}$, etc.): the welfare, measured as the aggregate ex ante gains from trade of all agents (and its derivatives).
- $x$: the investor’s market order size (net demand) at price $a$.

Additional notations in section 6:
- $a_j$: the $j$-th best ask price; $a_j \in \{p | p > v_0, p \in \mathcal{P}\}$.
- $j$: the $j$-th best (ask) price level.
- $\mathcal{P}$: the collection of all possible price levels in the order book with $\min\mathcal{P} = p_{\min} < \max\mathcal{P} = p_{\max}$.
- $[\theta_j]$ and $[\theta_j^+]$: the floor and the cap that bind the investor’s market order size at price level $j$ (see lemma 7).
- $x_j$: the investor’s market order size (net demand) at price level $j$.
- $y_{\leq j}$ (and $y_{<j}$, respectively): the cumulative depth up to and including (excluding) the $j$-th level.

Additional notations in appendix A:
- $\mu$: market maker $i$’s (expected) latency; $\mathbb{E}L_i = \mu$.
- $L_i$: market maker $i$’s random latency, exponentially distributed with mean $\mu_i$.

C Proofs

Lemma 1

Proof. First, note that by quasi-concavity of $\pi(\cdot)$ and the terminal conditions, there exists a $y^\equiv > 0$ such that $\dot{\pi}(y^\equiv) = 0$, i.e. the $y^\equiv$-th marginal order breaks even (in expectation). Second, no market maker will submit a limit order larger than $y^\equiv$ because the part exceeding $y^\equiv$ always loses (negative marginal profit). By reducing the order size to $y^\equiv$ strictly improves his total profit. This reduces each market maker’s strategy space to $[0, y^\equiv]$, and the best response correspondences implied by the first-order conditions can be summarized as a vector-valued function $f : [0, y^\equiv]^n \mapsto [0, y^\equiv]^n$, a nonempty, compact, and convex set. Note that the convexity of the value set of $f$ is implied by the continuity of the first-order conditions ($\pi(\cdot)$ is twice differentiable by assumption 1). Then, the existence of a Nash equilibrium follows Kakutani’s fixed point theorem. Finally, it is obvious that $q = 0$ is not an equilibrium because at least one market maker wants to deviate ($\dot{\pi}(0) > 0$, assumption 1). Therefore, at least one $q_i$ is strictly positive in equilibrium. \hfill $\square$

Lemma 2

Proof. Rearranging the zero-profit condition 6 to get $\pi(y(h)) = -\lambda(h; \delta)^{-1}\pi h \left(\min[k_\eta^\delta, \delta/\eta]q(h)\right) < 0$, where $y(h)$ is the true book depth defined in equation (3). The inequality follows due to the construction of $k_\eta^\delta$. This implies, by assumption 1, that $y(h) > y^\equiv$, i.e. there is liquidity overshoot. The above inequality holds for all $0 \leq h < \bar{h}$. At $h = \bar{h}$, by definition, the book stabilizes and hence there is no more liquidity overshoot. \hfill $\square$

Lemma 3

Proof. The lemma directly follows the analysis in section 4.2. \hfill $\square$
Lemma 4

Proof. First, fix $\delta$ and examine changes in $\eta$. Denote by $\bar{\bar{y}}(h)$ the stabilized part of the true depth (see equation 3) before the $h$-th revision. Then $\bar{\bar{y}}(h) = \sum_{j=0}^{h-1} \delta q(j)/\eta$, and

$$\frac{\partial \bar{\bar{y}}(h)}{\partial \eta} = -\frac{\delta}{\eta^2} \left[ \sum_{j=0}^{h-1} q(j) - \eta \sum_{j=0}^{h-1} \frac{\partial q(j)}{\partial \eta} \right] \approx -\frac{1}{\eta} \bar{\bar{y}}(h) \leq 0,$$

where the approximation is valid for small $\eta$. Hence, as the transmission latency reduces, the stabilized book depth before each revision (weakly) increases. Suppose $\bar{h} = h$. Then $\bar{\bar{y}}(h) \geq \bar{y}^\infty$ but $\bar{\bar{y}}(h-1) < \bar{y}^\infty$ and as $\eta$ drops, $\bar{\bar{y}}(h-1)$ keeps increasing until it reaches $\bar{\bar{y}}(h-1) = \bar{y}^\infty$. At $\bar{\bar{y}}(h-1) = \bar{y}^\infty$, after the stopping criterion (7) is satisfied by $\bar{h} = h-1$. Hence, $\bar{h}$ is an increasing right-continuous function in $\eta$. Note that at $\eta \to 0$, $\bar{h}$ reaches its minimum at $\bar{h} = 1$.

Similarly, fix $\eta$ and examine changes in $\delta$ next.

$$\frac{\partial \bar{\bar{y}}(h)}{\partial \delta} = \sum_{j=0}^{h-1} \frac{1}{\eta} q(j) + \delta \sum_{j=0}^{h-1} \frac{1}{\eta} \frac{\partial q(j)}{\partial \delta} \approx \frac{1}{\delta} \bar{\bar{y}}(h) \geq 0,$$

where the approximation is valid for small $\delta$. Hence, as the reaction latency increases, the stabilized book depth before each revision (weakly) increases, and in particular, if $\bar{h} = h$, then $\bar{\bar{y}}(h-1)$ keeps increasing, until it reaches $\bar{\bar{y}}(h-1) = \bar{y}^\infty$, satisfying the stopping criterion (7) at $\bar{h} = h-1$. That is, $\bar{h}$ is a decreasing left-continuous function in $\delta$. Note that if $\delta \to 0$, $\bar{h} \to \infty$. □

Lemma 5

Proof. Fix a limit order vector $q$ and a possible queue realization $k$. The expected profit of market maker $i$, given the queue, is $\pi \left( Q^i_q(k) + q_i \right) - \pi \left( Q^i_q(k) \right)$. Sum over all market makers’ expected profit to get

$$\sum_i \left[ \pi \left( Q^i_q(k) + q_i \right) - \pi \left( Q^i_q(k) \right) \right] = [\pi(q_{i(1)}) - \pi(0)] + [\pi(q_{i(1)} + q_{i(2)}) - \pi(q_{i(1)})] + \cdots + \left[ \pi \left( \sum_{k=1}^n q_{i(k)} \right) - \pi \left( \sum_{k=1}^{n-1} q_{i(k-1)} \right) \right] = \pi \left( \sum_i q_i \right).$$

In the equation, subscripts on $q$ means that $i(k)$ is the index of the agent at the $k$-th position of the queue. The last equality follows because all the in-between $\pi(\cdot)$ terms cancel. (Recall also that $\pi(0) = 0$.) The aggregate expected profit is queue-irrelevant. □

Lemma 6

Proof. That $w(y) > \pi(y)$ is trivial because by definition welfare accounts for both the certainty equivalent of the investor and the expected profit of market makers: $w(y) = \pi(y) + ce(y)$. Evaluate the difference between $w(y)$ and $\bar{\pi}(y)$ to get $\bar{w}(y) - \bar{\pi}(y) = \int_{[\theta] + py}^{\infty} (v_0 + \theta - \rho y - a) f(\theta) d\theta > 0$. The integration is positive because $\theta > [\theta] + py = a - v_0 + py$. Therefore, for all $y \geq 0$, $\bar{w}(y) > \bar{\pi}(y)$. By proposition 5, $\pi(y)$ exhibits quasi-concavity and hence, for all $0 \leq y \leq \bar{y}^\infty$, $\bar{w}(y) > \bar{\pi}(y) \geq 0$. Similarly, after some simplification, $\bar{w}(y) - \bar{\pi}(y) = -\rho \cdot (1 - F([\theta] + py)) < 0$ for all $y$. By proposition 5, $\bar{w}(y) < \bar{\pi}(y) < 0$ on $y \in [0, \bar{y}^\infty]$. Hence, $w(y)$ is concavely increasing on $[0, \bar{y}^\infty]$. □
Lemma 7

Proof. Solving the lower and the upper corners of equation (14), i.e. \( x_j(\{\theta_j\}) = 0 \) and \( x_j(\{\theta_j\}) = y_j \), gives the floor and the ceiling. (Note that \( x_j(\theta) \) is monotone increasing in \( \theta \).) □

Lemma 8

Proof. The result follows equation (A.1): \( \mathbb{P}_\mu(K) = \prod_{j=1}^n [\mu^{-1}_{i,j}] / \sum_{h=j}^n \mu^{-1}_{i,h} \) = \( \prod_{j=1}^n [c \mu^{-1}_{i,j}] / \sum_{h=j}^n c \mu^{-1}_{i,h} \) = \( \mathbb{P}_{c\mu}(K) \), which holds for all \( c > 0 \). □

Lemma 9

Proof. By the closed-form solution of equation (A.1), the log-likelihood of a queue, \( k \), is

\[
\ln \mathbb{P}(K = k) = \ln \mathbb{P}(L_{i(1)} < \cdots < L_{i(n)}) = \sum_{j=1}^n \ln \mu^{-1}_{i,j} - \sum_{j=1}^n \ln \left( \sum_{h=j}^n \mu^{-1}_{i,h} \right)
\]

Consider agent \( i \) with position \( p \) in the queue. Denote the sub-queue of all other agents by \( K_{-i} \). Then fixing \( p \in \{1, \ldots, n\} \) and a sub-queue realization \( k_{-i} \), the log-likelihood of such a queue is

\[
\ln \mathbb{P}(K_i = p, k_{-i}) = \ln \mu^{-1}_i + \sum_{j \neq p} \ln \mu^{-1}_{i,j} - \sum_{j=1}^p \ln \left( \mu^{-1}_i + \sum_{h=j, h \neq p} \mu^{-1}_{i,h} \right) - \sum_{j=p+1}^n \ln \left( \sum_{h=j}^n \mu^{-1}_{i,h} \right)
\]

The marginal effect of an increase in \( \mu^{-1}_i \) on the log-likelihood is

\[
\frac{\partial \ln \mathbb{P}(K_i = p, k_{-i})}{\partial (\mu^{-1}_i)} = \mu_i - \sum_{j=1}^p \left[ \mu^{-1}_i + \sum_{h=j, h \neq p} \mu^{-1}_{i,h} \right]^{-1}
\]

In particular,

\[
\frac{\partial \ln \mathbb{P}(K_i = 1, k_{-i})}{\partial (\mu^{-1}_i)} = \frac{\mu_i - 1}{\mu^{-1}_i + \sum_{j \neq i} \mu^{-1}_j} > 0;
\]

\[
\frac{\partial \ln \mathbb{P}(K_i = n, k_{-i})}{\partial (\mu^{-1}_i)} = -\sum_{j=1}^{n-1} \left[ \mu^{-1}_i + \sum_{h=j}^{n-1} \mu^{-1}_{i,h} \right]^{-1} < 0;
\]

and for \( 1 \leq p < n \), the difference \( \ln \mathbb{P}(K_i = p, k_{-i}) - \ln \mathbb{P}(K_i = p + 1, k_{-i}) = 1/\left( \sum_{h=p+1}^n \mu^{-1}_{i,h} \right) > 0 \). That is, the marginal effect is monotone decreasing, from positive to negative, along the queue position, \( p \), of agent \( i \). This monotonicity holds for any sub-queue realization \( k_{-i} \). Therefore, for \( c > 1 \), there exists a \( p^* \in \{1, \ldots, n-1\} \) such that \( \mathbb{P}_\mu(K_i \leq k) = \sum_{p \leq k} \mathbb{P}_\mu(K_i = p) < \sum_{p \leq k} \mathbb{P}_{c\mu}(K_i = p) = \mathbb{P}_{c\mu}(K_i \leq k) \) if and only if \( p \leq p^* \), i.e. \( \mathbb{P}_{c\mu}(K_i \leq k) \) is first-order stochastically dominated by \( \mathbb{P}_\mu(K_i \leq k) \). The proof for the other half of the lemma is similar. □

Proposition 1

Proof. Suppose the opposite, \( \sum_i q_i < y \), holds in equilibrium. Then by the quasi-concavity of \( \pi(\cdot) \), \( \pi(y) > 0 = \pi(y^\#) \), \( \forall 0 \leq y \leq \sum_i q_i \). Note that for any queue realization \( k \), \( \sum_i q_i \leq \sum_j q_j \) by construction.
and the substitution rate is bounded by \( \dot{\pi}(Q_i^\theta + q_i) > 0 \). As the first-order condition (2) holds in equilibrium (lemma 1 guarantees the interior solution), this leads to a contradiction that the expectation of the product of strictly positive numbers equates zero. (The utility function is assumed to be strictly increasing.) Thus, the assumed inequality cannot hold in equilibrium. Instead, \( \sum_i q_i \geq y^\pi \) and by quasi-concavity, \( \dot{\pi}(\sum_i q_i) \leq 0 = \dot{\pi}(y^\pi) \).

Clearly, the “if and only if” part holds true in the trivial case of \( n = 1 \). Consider \( n \geq 2 \) in what follows. First, the “only if” direction: Suppose there is an equilibrium with \( \sum_i q_i = y^\pi \) (the equality holds) but there is no \( j \in \mathcal{N} \) such that \( \mathbb{P}(K_j = 1) = 1 \). Then for at least some \( j \in \mathcal{N}, 0 < \mathbb{P}(K_j = 1) < 1 \) (at least someone might, but not almost surely, be the first in queue). Consider the first-order condition of such a market maker. A contradiction is derived from the first-order condition of such a market maker because the left-hand side might, but not almost surely, be the first in queue). Consider the first-order condition of such a market maker.

\[ \mathcal{J} \]

Proof. Differentiate the left-hand side of the first-order condition (2) with respect to \( q_j \) for some \( j \neq i \) to get \( \delta^2 \mathbb{E} \pi/(\partial q_i \partial q_j) = \mathbb{E} \left[ \mathbb{I}_{\{K_i \neq K_j\}} \cdot \dot{\pi}(Q_i^\theta + q_i) \right] \leq 0 \), by concavity of \( \pi(\cdot) \). This then establishes that \( q_i \) and \( q_j \) are strategic substitutes. Next, in equilibrium, the first-order condition (2) holds and it implies an implicit function of \( q_i = q_i^*(q_j) \). The substitution rate is the first order derivative of \( q_i \) with respect to \( q_j \), which by implicit function theorem is \( \partial q_i^*/\partial q_j = -\mathbb{E}[\mathbb{I}_{\{K_i \neq K_j\}}] \bar{\dot{\pi}}(Q_i^\theta + q_i)/\mathbb{E}[\dot{\pi}(Q_i^\theta + q_i)] \). The partial derivative is negative because of the concavity, i.e. \( \bar{\dot{\pi}}(\cdot) < 0 \). Note that there is at least one queue in which \( k_j > k_i \) and the indicator function becomes zero. Hence, the numerator is always strictly less negative than the denominator and the substitution rate is bounded by \(-1, 0\).

\[ \mathcal{K} \]

Proposition 2

\[ \mathcal{L} \]

Proof. The comparative static results can be shown through the implicit function of the zero-profit condition (6) where \( h = 0 \). Note that the zero-profit condition is essentially a first-order condition. Therefore, its partial derivative with respect to \( q(0) \) must be negative in equilibrium (as the second-order condition is necessary for optimality). It then remains to sign the partial derivatives with respect to \( \delta \) and to \( \eta \). The partial derivative with respect to \( \delta \) can be written as

\[-\frac{\dot{\lambda}(\delta)}{\lambda(\delta)^2} \pi(\min\{y^\pi, \delta q(0)/\eta\}) + \frac{1}{\delta \lambda(\delta)} \frac{\delta}{\eta} q(0) \left( \frac{\delta}{\eta} q(0) \right) \mathbb{I}_{\{y^\pi < \delta q(0) < y^\pi\}} \].

If \( \delta q(0)/\eta \geq y^\pi \), then the partial derivative is negative as \( \dot{\lambda}(\delta) = e^{\delta/\tau} / \tau > 0 \) by assumption 4 and hence proves \( \partial q(0)/\partial \delta < 0 \). If \( \delta q(0) < y^\pi \), then note that \( \gamma \pi(y) < \pi(y) \) for \( 0 \leq y \leq y^\pi \) because \( \pi(y) \) is concavely increasing by assumption 2. Then the above expression can be further evaluated as, with \( y := \delta q(0)/\eta \),

\[-\frac{1}{\lambda(\delta) \delta} \left[ \dot{\lambda}(\delta) \delta \pi(y) - \lambda(\delta) y \pi(y) \right] < -\frac{1}{\lambda(\delta) \delta} \pi(y) \left[ \dot{\lambda}(\delta) \delta - \lambda(\delta) \right] < 0,

where the last inequality follows because \( \lambda(\delta) \) is convexly increasing for all \( \delta > 0 \).
Evaluate next the partial derivative with respect to \( \eta \) to get
\[
-\frac{\delta}{\eta^2} \frac{1}{\lambda(\delta)} g(0) \tilde{\pi}(\frac{\delta}{\eta} q(0)) \mathbb{I}_{\{\frac{\delta}{\eta} q(0) < y\}}.
\]
which is clearly non-positive as for all \( 0 \leq y \leq y^\circ \), \( \tilde{\pi}(y) > 0 \). This completes the proof. \( \square \)

**Proposition 4**

**Proof.** Note that there are two components in the expression of \( t^e \). Consider first the case where changes in the latencies do not affect \( \bar{h} \). Then \( \partial t^e / \partial \delta = (\bar{h} + 1) \bar{h}/2 + 1 - \eta \partial k_{h-1}^*/\partial \delta \approx (\bar{h} + 1) \bar{h}/2 + 1 \leq 0 \), where the approximation is valid for small \( \eta \). Similarly, \( \partial t^e / \partial \eta = (\bar{h} + 1) - k_{h-1}^* - \eta \partial k_{h-1}^*/\partial \eta \approx (\bar{h} + 1) - k_{h-1}^* > 0 \), where the approximation is valid for small \( \eta \) and the inequality follows because \( k^* \) measures a fraction of the market maker, whose maximal mass is 1. Therefore, if there is no jump of \( \bar{h} \), \( t^e \) increases (decreases) in \( \eta \) (in \( \delta \), respectively).

Fixing \( \eta \), suppose \( \bar{h} \) jumps down from \( h \) to \( h - 1 \) as \( \delta \) increases to \( \delta + \epsilon \), where \( \epsilon \) is an infinitesimally small amount. Then by definition, the second component in \( t^e \) is zero because \( k_{h-1}^* = \delta/\eta \) at the jump. The difference in \( t^e \) is simply \( t^e(\delta, \eta) - t^e(\delta + \epsilon, \eta) = \eta \cdot (1 - \bar{h} \delta/\eta) \geq 0 \) because by construction \( (1 - \bar{h} \delta/\eta) \) is the mass of market makers who still have queuing uncertainty about their initial orders by round \( h \). This mass must be non-negative (otherwise the revision would have stopped already). Hence, as reaction latency \( \delta \) decreases, the order book stabilizes in a longer period of time.

Now fixing \( \delta \), suppose \( \bar{h} \) jumps up from \( h \) to \( h + 1 \) as \( \eta \) increases to \( \eta + \epsilon \), where \( \epsilon \) is an infinitesimally small amount. Then by definition, the second component in \( t^e \) is zero because \( k_{h-1}^* = \delta/\eta \) at the jump. The difference in \( t^e \) is simply \( t^e(\delta, \eta) - t^e(\delta, \eta + \epsilon) = -\eta \cdot (1 - (h + 1) \delta/\eta) \leq 0 \). The inequality follows by the same argument as above. Hence, as transmission latency \( \eta \) decreases, the order book stabilizes in a shorter period of time. \( \square \)

**Proposition 5**

**Proof.** This proof shows the more general deep-in-book version of the proposition in section 6. From equation (16), the second-order derivative of \( \pi_j(y) \) can be computed as
\[
\hat{\pi}_j(y) = -\rho \left( a_j - \nu(\lfloor \theta_j \rfloor + \rho y) \right) f(\lfloor \theta_j \rfloor + \rho y).
\]
Suppose there exists some \( y^\circ \) such that \( \hat{\pi}_j(y^\circ) = 0 \). Then \( 0 = \hat{\pi}_j(y^\circ) = \int_{\lfloor \theta_j \rfloor + \rho y^\circ}^{\infty} (a_j - \nu(\theta)) f(\theta)d\theta < (a_j - \nu(\lfloor \theta_j \rfloor + \rho y^\circ)) \left( 1 - F(\lfloor \theta_j \rfloor + \rho y^\circ) \right) \), or, \( a_j - \nu(\lfloor \theta_j \rfloor + \rho y) = -\hat{\pi}_j(y) > 0 \). The inequality follows because \( \nu(\cdot) \) is assumed, as in Biais, Martimort, and Rochet (2000), to be increasing. By theorem M.C.4 of Mas-Colell, Whinston, and Green (1995), this implies the quasi-concavity of \( \pi_j(y) \). Hence, assumption 1 holds.

Similarly, suppose at some \( \hat{y} \) \( \hat{\pi}_j(\hat{y}) = 0 \), which implies \( a_j - \nu(\lfloor \theta_j \rfloor + \rho \hat{y}) = 0 \) because \( \Theta \) has a continuous support and hence the density function is strictly positive on its support. Then it can be evaluated that the third-order derivative of \( \pi_j(y) \) at such a \( \hat{y} \) is \( \rho^2 \nu(\lfloor \theta_j \rfloor + \rho y) f(\lfloor \theta_j \rfloor + \rho y) > 0 \) (as \( \nu(\cdot) > 0 \)). This implies that \( \hat{\pi}_j(y) \) is quasi-convex.

To prove assumption 2, suppose there exists some \( y_\gamma^\circ \) such that \( \hat{\pi}_j(y_\gamma^\circ) = 0 \). Then by the quasi-convexity of \( \pi_j(\cdot) \), \( \pi_j(y) > 0 \) for all \( 0 \leq y < y_\gamma^\circ \). Then, \( 0 \leq \hat{\pi}_j(y) = \int_{\lfloor \theta_j \rfloor + \rho y}^{\infty} (a_j - \nu(\theta)) f(\theta)d\theta < (a_j - \nu(\lfloor \theta_j \rfloor + \rho y)) \left( 1 - F(\lfloor \theta_j \rfloor + \rho y) \right) = -\pi_j(y) f(\lfloor \theta_j \rfloor + \rho y)/(1 - F(\lfloor \theta_j \rfloor + \rho y)), \) where the inequality “<”
follows the monotonicity of $v(\theta)$ and the last equality follows the expression of $\pi_j(\cdot)$. Rearrange the above inequality to get

\begin{equation}
\pi_j(y) \leq -\pi_j(y) \frac{f(\|\theta_j\| + \rho y)}{1 - F(\|\theta_j\| + \rho y)} < 0
\end{equation}

because the hazard rate is always positive. This shows that $\pi_j(y) > 0$ is sufficient for $\pi_j(y)$ to be concave. □

**Proposition 6**

*Proof.* Immediately following lemma 6, the welfare maximizes at some $y^* > y^*$ because $\dot{w}(y^*) > 0$. □

**Proposition 7**

*Proof.* Define $y := q_1 + q_2$ as the aggregate book depth. In equilibrium, the first-order condition system (12) holds (the interior solution is guaranteed by lemma 1) and by chain-rule,

\[ \frac{\partial q_1}{\partial \alpha} = \frac{\dot{\pi}(y)}{\Delta} \left[ 1 - \frac{\alpha}{\alpha} \dot{\pi}(q_2) + 2 \dot{\pi}(y) \right] > 0 \]

\[ \frac{\partial q_2}{\partial \alpha} = -\frac{\dot{\pi}(y)}{\Delta} \left[ \frac{\alpha}{1 - \alpha} \dot{\pi}(q_1) + 2 \dot{\pi}(y) \right] < 0 \]

where $\Delta = \alpha \cdot (1 - \alpha) \dot{\pi}(q_1) \dot{\pi}(q_2) + (1 - \alpha)^2 \dot{\pi}(q_2) \dot{\pi}(y) + \alpha^2 \dot{\pi}(q_1) \dot{\pi}(y) > 0$ is the determinant of the Jacobian matrix. To see why the partial derivatives (and $\Delta$) are so signed, note that by lemma 1, $0 \leq y = q_1 + q_2 \leq 2 y^*$ in equilibrium. Then by condition (13) $\dot{\pi}(y) < 0$ because $\dot{\pi}(y)$ is quasi-convex and $\dot{\pi}(0) < 0$ as implied by proposition 5. (Recall that $\dot{\pi}(y) < 0$ in equilibrium because of liquidity overshoot.) The effect on aggregate book depth, $y$, is

\[ \frac{\partial y}{\partial \alpha} = \frac{\partial q_1}{\partial \alpha} + \frac{\partial q_2}{\partial \alpha} = -\frac{\dot{\pi}(y)}{\Delta} \left[ \frac{\alpha}{1 - \alpha} \dot{\pi}(q_1) - \frac{1 - \alpha}{\alpha} \dot{\pi}(q_2) \right]. \]

By symmetry, $\dot{\pi}(y) > 0$ at $\alpha = 1/2$. It remains to show that $\partial^2 y / \partial \alpha^2 < 0$ at $\alpha = 1/2$, which is true as $\frac{\partial}{\partial \alpha} \bigg|_{\alpha=1/2} = -4 \dot{\pi}(y) \dot{\pi}(q_1)/\Delta < 0$ (note that $q_1 = q_2$ at $\alpha = 1/2$ by symmetry), making $\alpha = 1/2$ a local maximum. □

**Proposition 8**

*Proof.* The proof goes in several steps. Step one shows how to derive the first-order condition (18). Note that for $h \geq 1$, $\partial \pi_{j+h}(y)/\partial q_{i,j} = -\int_{[0,1]} f(\theta) \left( a_{j+h} - v(\theta) \right) d\theta = \pi_{j+h}(y) - \pi_{j+h}(0)$. Hence,

\begin{equation}
\frac{\partial}{\partial q_{i,j}} \left( \pi_{j+h}(y + q) - \pi_{j+h}(y) \right) = \pi_{j+h}(y + q) - \pi_{j+h}(y).
\end{equation}

The Bellman equation for price level $j + 1$ (see the $j$-th Bellman equation in equation 17) should hold in equilibrium: $u_{i,j+1} = \mathbb{E} \left[ \pi_{j+1} \left( Q_{i,j+1}^q + q_{i,j+1} \right) - \pi_{j+1} \left( Q_{i,j+1}^q \right) \right] + u_{i,j+2}$. Optimality of the right-hand side requires the first order condition to hold:

\begin{equation}
\mathbb{E} \pi_{j+1} \left( Q_{i,j+1}^q + q_{i,j+1} \right) + \frac{\partial u_{i,j+2}}{\partial q_{i,j+1}} = 0.
\end{equation}
Also, differentiate both sides of the Bellman equation with respect to \( q_{i,j} \) to get
\[
\frac{\partial u_{i,j+1}}{\partial q_{i,j}} = \frac{\partial}{\partial q_{i,j}} \mathbb{E} \left[ \pi_{j+1} (Q^e_{i,j+1} + q_{i,j+1}) - \pi_{j+1} (Q^e_{i,j+1}) \right] + \frac{\partial u_{i,j+2}}{\partial q_{i,j+1}} = -\mathbb{E} \hat{\pi}_{j+1} (Q^e_{i,j+1}),
\]
where the second equality follows by substituting equation (C.2) with \( h = 1, y = Q^e_{i,j+1}, \) and \( q = q_{i,j+1}, \) and equation (C.3). Finally, note that optimality of the Bellman equation (17) gives \( \mathbb{E} \hat{\pi}_{j} (Q^e_{i,j} + q_{i,j}) + \partial u_{i,j+1} / \partial q_{i,j} \). Substitute the above expression for \( \partial u_{i,j+1} / \partial q_{i,j} \) into the optimality, and the result is the first-order condition given in equation (18).

The second step derives an important properties about the expected profit function \( \pi_{j}(y) \). The quasi-concavity of \( \pi_{j}(y) \) and the quasi-convexity of \( \hat{\pi}_{j}(y) \) together imply the first property that if \( \hat{\pi}_{j}(y) \leq 0, \) then \( \pi_{j}(y) \leq 0 \) for all \( y \geq 0 \). To see this, note that if the opposite is true, i.e. if there exists some \( y \) such that \( \pi_{j}(y) > 0 \), then by continuity \( \pi_{j}(y) \) must cross the horizontal axis at least once from below to above. The crossing point implies a local minimum of \( \pi_{j}(y) \), contradicting the quasi-concavity of \( \pi_{j}(y) \).

With the above results, the last step is to suppose \( \hat{\pi}_{j}(y_{j}) \geq 0 \), where \( y_{j} := \sum q_{i,j} \), and show this leads to a contradiction. Note that \( (Q^e_{i,j} + q_{i,j}) \leq y_{j} \) in equilibrium because there is non-zero probability for market maker \( i \)'s order not to queue in the last position. Given that \( \hat{\pi}_{j}(y_{j}) \geq 0 \), therefore, \( \mathbb{E} \hat{\pi}_{j} (Q^e_{i,j} + q_{i,j}) \leq 0 \) by the inequality (C.1). Differentiate the first-order condition (18) on both sides with respect to \( q_{i,j} \):
\[
\mathbb{E} \hat{\pi}_{j} (Q^e_{i,j} + q_{i,j}) + \mathbb{E} \hat{\pi}_{j+1} (Q^e_{i,j+1}) = 0, \text{ or } \mathbb{E} \hat{\pi}_{j+1} (Q^e_{i,j+1}) = -\mathbb{E} \hat{\pi}_{j} (Q^e_{i,j} + q_{i,j}) > 0,
\]
where the inequality follows the above argument. Then a contradiction follows:
\[
0 < \mathbb{E} \hat{\pi}_{j+1} (Q^e_{i,j+1}) < -\mathbb{E} \left[ \pi_{j+1} (Q^e_{i,j+1}) \right] \frac{f([\theta_{j}] + Q^e_{i,j+1})}{1 - F([\theta_{j}] + Q^e_{i,j+1})} \frac{f([\theta_{j}] \rho)}{1 - F([\theta_{j}] \rho)} < 0
\]

The second inequality follows inequality (C.1). The third inequality follows the increasing hazard rate of \( \Theta \). The last inequality follows the first-order condition (18) that \( \mathbb{E} \hat{\pi}_{j+1} (Q^e_{i,j+1}) = \mathbb{E} \hat{\pi}_{j} (Q^e_{i,j} + q_{i,j}) \), which is positive because \( (Q^e_{i,j} + q_{i,j}) \leq y_{j} \) and because of the quasi-concavity of \( \pi_{j}(\cdot) \). To conclude, this contradiction refutes the assumption that \( \pi_{j}(y_{j}) \geq 0 \) in equilibrium. Therefore, \( \hat{\pi}_{j}(y_{j}) < 0 \), i.e. at each price level there is liquidity overshoot. Therefore, \( \hat{\pi}_{j}(y_{j}) < 0 \).

**Corollary 1**

**Proof.** Rewrite \( \partial y / \partial \alpha \) as
\[
\frac{\partial y}{\partial \alpha} = -\frac{\pi(y)}{\Delta} \hat{\pi}(q_{1}) \frac{\alpha}{1 - \alpha} \left( \hat{\pi}(q_{1}) - \hat{\pi}(q_{2}) \right)
\]
by noting \( \hat{\pi}(q_{1}) / \hat{\pi}(q_{2}) = (1 - \alpha)^2 / \alpha^2 \). Because \( \partial q_{1} / \partial \alpha > 0 \) and \( \partial q_{2} / \partial \alpha < 0 \) (see the proof of proposition 7), for \( \alpha \geq 1/2, q_{1} \geq q_{2} \) (that is, the faster market maker supplies more). Therefore, it suffices to sign the derivative of \( \hat{\pi}(x) / \hat{\pi}(x) \) on \( x \in [0, y_{\alpha}] \): If \( \hat{\pi}(x) / \hat{\pi}(x) \) is decreasing, then \( \partial y / \partial \alpha \leq 0 \) and \( y \) reaches its global maximum at \( \alpha = 1/2 \). Note that
\[
\frac{\partial}{\partial x} \left( \frac{\hat{\pi}(x)}{\hat{\pi}(x)} \right) = \frac{1}{\hat{\pi}(x)^2} \left( \hat{\pi}(x) \hat{\pi}(x) - \hat{\pi}(x)^2 \right)
\]
is negative if \( \hat{\pi}(x) \leq 0 \) \( \hat{\pi}(x) > 0 \) for \( x \in [0, y_{\alpha}] \) by quasi-concavity of \( \pi_{j}(\cdot) \). That is, concavity of \( \hat{\pi}(x) \) on \( [0, y_{\alpha}] \) is a sufficient condition for the result.
Consider next the case where \( \hat{\pi}(x) \) is not concave but only the first three order effects matter:

\[
\hat{\pi}(x) = c_1 \cdot (x - y^\circ) + c_2 \cdot (x - y^\circ)^2 + O((x - y^\circ)^3) \approx c_1 \cdot (x - y^\circ) + c_2 \cdot (x - y^\circ)^2.
\]

Then it can be derived immediately that

\[
\frac{\partial}{\partial x} \left( \frac{\hat{\pi}(x)}{\pi(x)} \right) = \frac{1}{\hat{\pi}(x)^2} \left( \hat{\pi}(x) \hat{\pi}(x) - \hat{\pi}(x)^2 \right) = -\frac{1}{\pi(x)^2} \left( c_2^2 \cdot (x - y^\circ)^2 + (c_1 + c_2 \cdot (x - y^\circ))^2 \right) < 0.
\]

Hence, when the above approximation holds, \( \partial y / \partial \alpha \) is negative on \( \alpha \geq 1/2 \) and, by symmetry, positive on \( \alpha \leq 1/2 \). At \( \alpha = 1/2 \), \( y \) has the unique maximum.

\( \square \)

**Corollary 2**

**Proof.** The proof directly follows corollary 1, which implies the equilibrium book depth \( y \) is a one-to-one mapping from the effective queue distribution parameter, \( \alpha_R \), which in turn is a one-to-one mapping from \( \beta \), which governs the queue randomizer. Therefore, by choosing \( \beta \) (hence also \( R \)), the equilibrium depth can be adjusted to maximize welfare, which is a function in \( y \): \( w(y) = w(y(\alpha_R(\beta))) \). (Note that the optimal level \( \beta \) might be cornered.)

\( \square \)

**Corollary 3**

**Proof.** Suppose the equilibrium depth at some price level \( a_j \) is \( y_j > 0 \). By proposition 8, \( \hat{\pi}(y_j) = \int_{[\theta_j]^{+}+\rho y_j} (a_j - v(\theta)) f(\theta) d\theta < 0 \). Then the expected marginal profit of the first unit of limit orders at the next possible price, \( a_j + \rho \), is \( \int_{[\theta_{j+}]^{+}+\rho y_j} (a_j + \rho - v(\theta)) f(\theta) d\theta \), which is negative for sufficiently small \( \rho \). Then by property 1) developed in the proof of proposition 8, the expected marginal profit at price \( a_j + \rho \) is always negative for all depth \( y \geq 0 \). Hence, no market maker is willing to post any order at this price level, leaving it a hole in the book.

\( \square \)

**Corollary 4**

**Proof.** In this proof, the superscript “\(*\)” indicates “with queuing uncertainty”, while the superscript “\( \equiv \)” indicates “under the break-even condition of Glosten (1994)”. Consider the expected marginal profit functions at price level \( j \): \( \hat{\pi}_j^\equiv(y) = \int_{[\theta_j]^{+}+\rho y_j} (a_j^\equiv - v(\theta)) f(\theta) d\theta \) and \( \hat{\pi}_j^\neq(y) = \int_{[\theta_j]^{+}+\rho y_j} (a_j^\neq - v(\theta)) f(\theta) d\theta \). In particular, \( [\theta_j]^+ = a_j^+ - v_0 + \rho y_j^+ \) and \( [\theta_j]^\equiv = a_j^\equiv - v_0 + \rho y_j^\equiv \). Decompose \( \hat{\pi}_j(y) \) as

\[
\hat{\pi}_j^\equiv(y) = \int_{[\theta_j]^+\neq+\rho y_j} (a_j^\equiv - v(\theta)) f(\theta) d\theta + (a_j^\equiv - a_j^\neq) \left( 1 - F([\theta_j]^+ + \rho y_j^\equiv) \right) > 0.
\]

When there is no holes in the order book, \( a_j^\equiv = a_j^\neq \), and the following inequality follows at equilibrium:

\[
\hat{\pi}_j^\neq \left( y_j^+ + (a_j^\neq - a_j^\equiv) / \rho + (y_j^\equiv - y_j^\neq) \right) < \hat{\pi}_j^\equiv(y_j^+) < 0 = \hat{\pi}_j(y_j^\equiv).
\]

By quasi-concavity of \( \pi_j(\cdot) \), therefore, \( y_j^+ + (y_j^\equiv - y_j^\neq) > y_j^\equiv \), which simplifies to \( y_j^\equiv > y_j^\equiv \).

\( \square \)
References


List of Figures

1. Game structure comparison ................................................. 2
2. Equilibrium illustration with \( n = 2 \) market makers ..................... 12
3. Time line of the dynamic game ............................................. 16
4. Order book depth dynamics with varying latency ....................... 23
5. Welfare as a function of liquidity supply ................................ 30
6. Queuing uncertainty and welfare .......................................... 35
7. The investor’s optimal market order size ................................ 37