# *Maker-Taker Fees and Informed Trading in a Low-Latency Limit Order Market*[*]

Michael Brolley[†] Katya Malinova[‡]
University of Toronto University of Toronto

October 18, 2012

— *preliminary* —

### Abstract

We model a financial market with investors that trade for informational and liquidity reasons in a limit order book that is monitored by low-latency liquidity providers. We apply our model to study the impact of the commonplace, but controversial maker-taker fee system, which imposes differential trading fees on liquidity providers (makers) and removers (takers). In our benchmark setting, the maker-taker fees are passed through to all traders, and only the net fee (the amount that the exchange receives) has an economic impact, consistent with the previous literature. When instead some investors pay only the average exchange fee, through a flat fee per transaction, a disparity in liquidity provision incentives between investors and low-latency liquidity providers arises, and the split between maker and taker fees matters. A decrease in the maker fee increases trading volume, lowers trading costs, but decreases market participation by investors. Finally, we find that the common industry practice of subsidizing liquidity provision through a negative maker fee is welfare enhancing.

[*] Financial support from the Social Sciences and Humanities Research Council is gratefully acknowledged.

[†] Email: michael.brolley@utoronto.ca; web: http://individual.utoronto.ca/brolleym/.

[‡] Email: katya.malinova@utoronto.ca; web: http://individual.utoronto.ca/kmalinova/.

Equity trading around the world is highly automated. Exchanges maintain limit order books, where orders to trade pre-specified quantities at pre-specified prices are arranged in a queue, according to a set of priority rules.[1] A trade occurs when an arriving trader finds the terms of limit orders at the top of the queue sufficiently attractive and posts a marketable order that executes against these posted limit orders.

To improve the trading terms, or liquidity, offered in their limit order books, many exchanges incentivize traders who provide, or "make" liquidity. Specifically, trading venues pay a rebate to submitters of executed limit orders, and they finance these rebates by levying higher fees to remove, or "take" liquidity on submitters of marketable orders. This practice of levying different trading fees for liquidity provision and removal is referred to as "maker-taker" pricing. Moreover, with the rise of algorithmic trading, exchanges have adopted technology that offers extremely high-speed, or "low-latency" market data transmission, in order to appeal to speed-sensitive participants. The rebates, along with the increased speed of trading systems, has given rise to "a new type of professional liquidity provider": proprietary trading firms that "take advantage of low-latency systems" and provide liquidity electronically.[2]

The role of maker-taker pricing and the new low-latency computerized traders remains controversial. Proponents maintain that the new trading environment benefits all market participants through increased competition. Opponents argue that the increased competition for liquidity provision makes it difficult for long-term investors to trade via limit orders and that it compels them to trade with more expensive marketable orders.[3] As a practical matter, however, many long-term investors do not pay taker fees directly and do not receive maker rebates but instead pay a flat fee per trade to their broker.

[1] Most exchanges sort limit orders first by price, and then by the time of arrival, maintaining a so-called price-time priority.

[2] SEC Concept Release on Market Structure, Securities and Exchange Commission (2010)

[3] See, e.g., GETCO's comments on maker-taker fees in options markets to SEC (available at http: //www.getcollc.com/images/uploads/getco_comment_090208.pdf), in favor, and TD Securities' comments on IIROC 11-0225 (www.iiroc.ca), Alpha Trading Systems' September 2010 Newsletter (http://www.alphatradingsystems.ca/, against.

This practice of levying flat fees has recently become an additional issue in the debate, after an industry study argued that the maker-taker pricing in its current form distorts trading incentives and causes losses to long-term investors.[4]

In this paper, we develop a dynamic model to analyze trader behavior in the presence of low-latency liquidity providers. We then employ our model to study the impact of maker-taker fees on liquidity, trading volume, and market participation. We address the effect of maker-taker fee implementation, by comparing investor trading incentives in a benchmark setting, where all market participants incur maker and taker fees, to a setting where long-term investors pay a flat fee per trade, which equals the average of per trade fees charged by the exchange on long-term investor trades. Our model illustrates that maker-taker pricing in its current form, where brokers do not pass maker rebates and taker fees to long-term investors on a trade-by-trade basis, may have a detrimental effect on investor market participation; however, it leads to an improvement in observed trading costs and to an increase in trading volume.[5]

In our model, risk-neutral traders sequentially enter the market for a risky security to trade on private information and for liquidity reasons. Traders may submit an order to buy or sell one unit (one round lot) upon entering and only then, or may abstain from trading. Trading in our model is organized via limit order book, and traders can choose between submitting a limit order or a market order. Additionally, some traders trade solely for the purpose of providing liquidity and they only submit limit orders. These professional liquidity providers permanently monitor the market, compete in prices, (in the sense of Bertrand competition) and posses a speed advantage that allows them to react to changes in the limit order book faster than other market participants. We

[4] On May 10, 2012, Senator Schumer called on the Securities and Exchanges Commission to mandate that taker fees and maker rebates are passed through to investors. Commenting on this, Larry Tabb of market-research firm Tabb Group raised concerns that passing through the exchange fees may “disadvantage investors because they’re generally takers of liquidity” (*The Wall Street Journal*, “Schumer to SEC: Fix ’Maker-Taker’ Fees”.).

[5] Addressing broker-dealer trading incentives and possible agency costs stemming from conflicts of interest between investors and their brokers is outside the scope of this paper, and we may thus overstate the benefits of the maker-taker pricing model.

refer to them as *low-latency liquidity providers*, and we refer to traders that trade for liquidity or informational needs as *investors*. Upon entering the market, an investor observes the history of past transactions and quote revisions, as well as the current state of the limit order book. Each investor has a private valuation for the security stemming from liquidity needs, and additionally, receives private information about the value of the security. Low-latency liquidity providers are uninformed and have no liquidity needs.

Our setup captures the low-latency liquidity providers' speed advantage in interpreting market data, such as trades and quotes. The speed advantage comes at a cost, however, and low-latency liquidity providers are arguably at a disadvantage (relative to humans or sophisticated algorithms) when processing more complex information, such as news reports. We capture this difference in information processing skills by allowing investors an informational advantage with respect to the security's fundamental value.

The presence of low-latency liquidity providers lends tractability to our setup: competition among them induces all limit order submitters to offer competitive prices and thus pins down limit order prices in equilibrium. With competitive pricing, a limit order price is the expected value of the security, conditional on submission and on execution of this limit order. The price of, say, a buy limit order is then determined deterministically; loosely, by the average information of traders who submit buy limit orders and the average information of traders who would submit sell market orders in the next period.

In equilibrium, a trader's behavior is governed by their aggregate valuation, which is the sum of his private valuation of the security and his expected value of the security. Traders with extreme aggregate valuations optimally choose to submit market orders, traders with moderate valuations submit limit orders, and traders with aggregate valuations close to the public expectation of the security's value abstain from trading.

To analyze the impact of maker-taker fees and their implementation, we compare two settings. In the benchmark setting, low-latency liquidity providers and investors both incur maker and taker fees for executed limit and market orders, respectively. In the second setting, only low-latency liquidity providers access the exchange directly and pay (possibly negative) maker fees. Investors, on the other hand, submit their orders via a broker, who charges investors a flat fee per trade and who then pays taker fees directly to and collects maker rebates directly from the exchange. We assume that brokers act competitively, in the sense that the flat fee per trade is the average exchange fee that a broker incurs per trade when executing trades on behalf of investors.

We specifically focus on the impact of the split of the total exchange fee into a maker fee and a taker fee. Consistent with the previous literature (see Angel, Harris, and Spatt (2011) and Colliard and Foucault (2012)), when maker-taker fees are passed through, the split does not play an economically meaningful role in our model, because any decrease in the maker fee is passed to the takers through a lower bid-ask spread, exactly offsetting an increase in the taker fee.

When only low-latency liquidity providers pay maker fees, a decrease in a maker fee will, ceteris paribus, lower the bid-ask spread and therefore induce investors who were previously indifferent between a market and a limit order to trade with market orders (since an investor's trading cost consists, loosely, of the bid-ask spread and the flat fee levied by their broker). Consequently, the probability of a market order submission increases, and so does the trading volume (low-latency liquidity providers ensure that the limit order book is always full in our setup).

We find numerically that, for a fixed total exchange fee, as the maker fee decreases and the taker fee increases, investors submit more market orders, fewer limit orders, and further, more investors choose to abstain from trading. This leads to brokers paying taker fees more frequently and consequently charging investors a higher flat fee. The increase in the flat fee is more than offset, however, by the decline in the bid-ask spread, and investors' overall trading costs decline. The marginal submitter of a market order requires weaker information, and thus the price impact of a trade declines.

When the maker fee is sufficiently small (and negative, in our setup) and the taker fee is sufficiently large, in equilibrium, investors choose to trade exclusively with market orders. The average fee charged by brokerages then equals the taker fee, and any changes in it are exactly offset by changes in the quoted bid-ask spread. In particular, any decrease in the maker fee that is financed by an increase in the taker fee then leads to a decline in the quoted spreads, but yields no economically meaningful implications.

To analyze the impact of maker-taker fees on welfare, we follow Bessembinder, Hao, and Lemmon (2012) and define a social welfare measure to reflect allocative efficiency. Specifically, with each trade, the social gains from trade increase by the difference between the buyer's and the seller's private valuations, net of differences in trading fees, and we define the social welfare to be the expected social gains per period. When the maker fee declines (and the taker fee increases by the same amount), two changes happen. First, some investors switch from submitting limit orders to trading with market orders, increasing the execution probability of their own order (to certainty), and also increasing the execution probability of a limit order, the so-called fill rate, for the remainder of the limit order submitters. Second, some investors switch from submitting limit orders to abstaining from trade, failing to realize any potential gains from trade.

We find numerically that the benefit of an increased fill rate to investors who remain in the market exceeds the loss of potential gains from trade to investors who choose to leave the market, and that welfare increases as the maker fee declines. The prevalent industry practice of setting a negative maker fee (i.e., a positive maker rebate) is thus socially optimal.

Our paper is most closely related to Colliard and Foucault (2012) and Foucault, Kadan, and Kandel (2012), who theoretically analyze the impact of maker-taker fees. Colliard and Foucault (2012) study trader behavior in a model where symmetrically informed traders choose between limit and market orders. They show that, absent frictions, the split between maker and taker fees has no economic impact, and they focus on the impact of the total fee charged by an exchange. Foucault, Kadan, and Kandel

(2012) argue that in the presence of a minimum tick size, limit order book prices may not adjust sufficiently to compensate traders for changes in the split between maker and taker fees. They then show that exchanges may use maker-taker pricing to balance supply and demand of liquidity, when traders exogenously act as makers or takers. Skjeltorp, Sojli, and Tham (2012) support theoretical predictions of Foucault, Kadan, and Kandel (2012) empirically, using exogenous changes in maker-taker fee structure and a technological shock for liquidity takers. They find that a decrease in taker fees increases takers' response speed to changes in liquidity, and they further identify positive liquidity externalities between makers and takers. We contribute to this strand of literature by analyzing a different friction, namely, where maker-taker fees are only passed through to investors on average. Our predictions on spreads, price impact, and volume are supported empirically by Malinova and Park (2011), who study the impact of the introduction of maker rebates on the Toronto Stock Exchange.

The maker-taker pricing model is related to the payment for order flow model, see, e.g., Kandel and Marx (1999), Battalio and Holden (2001), or Parlour and Rajan (2003), in the sense that both systems aim to incentivize order flow. Most recently, Battalio, Shkilko, and Van Ness (2012) and Anand, McCormick, and Serban (2012) empirically compare trading costs under the maker-taker pricing with those under the payment for order flow structure in the U.S. options markets, where the two fee models coexist. Degryse, Achter, and Wuyts (2009) theoretically study the impact of clearing and settlement fees on liquidity and welfare.

Our paper contributes to the broader theoretical literature on limit order markets, see e.g., Glosten (1994), Parlour (1998), Foucault (1999), Foucault, Kadan, and Kandel (2005), Goettler, Parlour, and Rajan (2005), and Rosu (2009), for limit order books with uninformed liquidity provision, and Kaniel and Liu (2006), Goettler, Parlour, and Rajan (2009), and Rosu (2011), for informed liquidity provision.[6] Our analysis of investor

[6] See also the survey by Parlour and Seppi (2008) for further related papers.

behavior in the presence of low-latency liquidity providers complements theoretical literature that focuses on the trading strategies of low-latency traders, see e.g., Biais, Foucault, and Moinas (2012), McInish and Upson (2012), and Hoffmann (2012).

Finally, the role of low-latency traders as competitive liquidity providers is supported empirically by e.g., Hasbrouck and Saar (2011), Hendershott, Jones, and Menkveld (2011), Hendershott and Riordan (2012), and Jovanovic and Menkveld (2011).

# 1 The Model

We model a financial market where risk-neutral investors enter the market sequentially to trade a single risky security for informational and liquidity reasons (as in Glosten and Milgrom (1985)). Trading is conducted via limit order book. Investors choose between posting a limit order to trade at pre-specified prices and submitting a market order to trade immediately with a previously posted limit order. Additionally, we assume the presence of low-latency liquidity providers, who choose to act as market makers, and to only submit limit orders. These traders possess a speed advantage that allows them to react to changes in the limit order book faster than other market participants. We assume that they are uninformed and that they have no liquidity needs. Low-latency liquidity providers compete in the sense of Bertrand competition, are continuously present in the market, and they ensure that the limit order book is always full.

**Security.** There is a single risky security with an unknown liquidation value. This value follows a random walk, and at each period $t$ experiences an innovation $\delta_t$. The fundamental value at period $t$ is given by

$$V_t = \sum_{\tau \leq t} \delta_\tau \tag{1}$$

Innovations $\delta_t$ are identically and independently distributed, according to density function $\bar{g}$ on $[-1, 1]$, which is symmetric around zero. We focus on intraday trading, and we assume that extreme innovations to the security's fundamental value are less likely than innovations that are close to 0, (i.e., that $\bar{g}'(\cdot) \leq 0$ on $[0, 1]$).

**Investors.** There is a continuum of risk-neutral investors. At each period $t$, a single investor randomly arrives at the market. Upon entering the market, the investor is endowed with liquidity needs, which we quantify by assigning the investor a private value for the security, denoted by $y_t$, uniformly distributed on $[-1, 1]$. Furthermore, the investor learns the period $t$ innovation to the fundamental value, $\delta_t$.[7]

**Investor Actions.** An investor can submit an order upon arrival and only then. He can buy or sell a single unit (round lot) of the risky security, or abstain from trading.[8] If the investor chooses to buy, he either submits a market order and trades with an existing order at the previously posted ask price $\mathsf{ask}_t$ in period $t$, or he posts a limit buy order at the bid price $\mathsf{bid}_{t+1}$ in period $t$, for execution in period $t + 1$. Similarly for the decision to sell. Limit orders that are submitted in period $t$ and that do not execute in period $t + 1$ are automatically cancelled. An investor may submit at most one order, and upon the order's execution or cancellation the investor leaves the market forever.

**Low-Latency Liquidity Providers.** There is continuum of low-latency liquidity providers who are always present in the market. They hold a speed advantage in reacting to changes in the limit order book. These traders act as market makers and post limit orders in response to changes in the limit order book. They compete in prices in the sense of Bertrand competition. Low-latency liquidity providers are risk-neutral, they do not receive any information about the security's fundamental value, and they do not have liquidity needs.

**The Limit Order Book.** Trading is organized via limit order book, which is comprised of limit orders. Limit orders last for one period. Arguably, this simplifying

[7] Assuming that traders have liquidity needs is common practice in the literature on trading with asymmetric information, to avoid the no-trade result of Milgrom and Stokey (1982). We also solved for an equilibrium, assuming that only a fraction of traders become informed, with qualitatively similar results.

[8] We will refer to investors in the male form, and we will refer to the low-latency liquidity providers in the female form.

assumption is particularly realistic in presence of low-latency traders, as slower investors may fear that their orders become stale and will be “picked off” by the low-latency traders. Low-latency liquidity providers ensure that the limit order book is always “full” by submitting a limit order when there is no standing limit order on the buy or the sell side. The limit order book thus always contains one buy limit order and one sell limit order, upon arrival of an investor in period $t$. A trade occurs in period $t$ when the investor that arrives in period $t$ chooses to submit a market order.

**Trading Fees.** The limit order book is maintained by an exchange that charges fees for executing orders. These fees depend on the order type (market or limit), and they may depend on the trader type; the trading fees do not depend on whether an order is a “buy” or a “sell” and they are independent of $t$. We discuss further details in Section 3.

**Public Information.** Investors and low-latency liquidity providers observe the history of transactions as well as limit order submissions and cancellations. We denote the history of trades and quotes up to (but not including) period $t$ by $H_t$. The structure of the model is common knowledge among all market participants, but an investor’s liquidity needs and his knowledge of an innovation to the fundamental value are private.

**Low-Latency Liquidity Provider Information.** Low-latency liquidity providers are able to detect whether a newly posted limit order stems from an investor with liquidity and informational needs or from other low-latency liquidity providers. This assumption ensures that the model is tractable. We believe that it is consistent with reality, because low-latency traders are allegedly good at identifying, for instance, larger institutional orders. Further, within our model, low-latency liquidity providers react virtually instantaneously to changes in the limit order book, whereas investors who trade for liquidity and informational reasons arrive at discrete time intervals — consequently, limit orders that are posted by low-latency liquidity providers are identified by the reaction time. Finally, from a technical perspective, this assumption is equivalent to assuming presence of a single low-latency liquidity provider who chooses to act competitively.

**Timing of Actions.** We model intraday trading. Periods are measured in discrete units (which we denote by $t$) with no specific beginning or end. Each period marks the arrival of an investor. At the beginning of any period $t$, the limit order book is full in the sense that it contains one buy limit order and one sell limit order. At each period $t$, an investor enters the market, observes the transaction and quote history $H_t$, his liquidity needs measured by his private valuation $y_t$, and the innovation $\delta_t$ to the security's value. This investor posts a limit or a market order, or abstains from trading.

When a market order is posted, it executes against a limit order that was posted at period $t-1$, and the investor leaves the market forever. The limit order book immediately reacts to the information contained in the period $t$ market order and the low-latency liquidity providers post limit orders to buy and sell.

When a limit order is posted at period $t$, this order remains in the market until the period $t+1$ investor makes his trading choice. This limit order possibly interacts with the period $t+1$ investor's market order. As with market orders, the limit order book reacts to the information contained in the period $t$ limit order, with a low-latency liquidity provider posting a limit order on the opposite side of the book.

**Investor Payoffs.** The payoff to an investor who buys one unit of the security at period $t$ is given by the difference between the security's fundamental value at period $t$, $V_t$, and the price that the investor paid for this unit; similarly for a sell decision. We normalize the payoff to a non-executed order to 0. Investors are risk neutral, and they aim to maximize their expected payoffs. The period $t$ investor with private valuation $y_t$ has the following expected payoffs to submitting, respectively, a market buy order to trade immediately at the prevailing ask price $\mathsf{ask}_t$ and a limit buy order at price $\mathsf{bid}_{t+1}$:

$$\pi_t^{\text{MB}}(y_t, \delta_t) = y_t + \mathsf{E}[V_t \mid \delta_t, H_t] - \mathsf{ask}_t - \text{fee}_{\text{inv}}^{M}, \tag{2}$$

$$\begin{aligned}\pi_t^{\text{LB}}(y_t, \delta_t, \mathsf{bid}_{t+1}) &= \mathsf{Pr}(\text{MS}_{t+1}(\mathsf{bid}_{t+1}) \mid \delta_t, H_t) \times (y_t + \\ &\quad + \mathsf{E}[V_{t+1} \mid \big(\delta_t, H_t, \text{MS}_{t+1}(\mathsf{bid}_{t+1})] - \mathsf{bid}_{t+1} - \text{fee}_{\text{inv}}^{L}\big)\,,\end{aligned} \tag{3}$$

where $\mathrm{MS}_{t+1}(\mathsf{bid}_{t+1})$ represents the period $t+1$ investor's decision to submit a market order to sell at price $\mathsf{bid}_{t+1}$ (this decision is further conditional on the additional information available to the period $t+1$ investor); $\mathrm{fee}^{M}_{\mathrm{inv}}$ and $\mathrm{fee}^{L}_{\mathrm{inv}}$ denote the fees incurred by investors when trading with market and limit orders, respectively. An investor's payoff to submitting a limit order at period $t$ accounts for the fact that a limit order submitted at period $t$ either executes or is cancelled at period $t+1$. We focus on the intraday trading, and we assume no discounting. Payoffs to sell orders are defined analogously.

**Low-Latency Liquidity Provider Payoffs.** A low-latency trader observes the period $t$ investor's action before posting her period $t$ limit order. Moreover, she will post a limit buy order at period $t$ only if the period $t$ investor does not post a buy limit order.[9] Denoting by $\mathrm{fee}^{L}_{\mathrm{LLT}}$ the trading fee incurred by a low-latency trader when her limit order is executed, a low-latency trader at period $t$ has the following payoff to submitting a limit buy order at price $\mathsf{bid}_{t+1}$ is given by

$$\begin{aligned}\pi^{\mathrm{LB}}_{t,\mathrm{LLT}}(\mathsf{bid}_{t+1}) \quad = \quad & \mathsf{Pr}(\mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}) \mid \text{investor action at } t, H_t) \qquad (4)\\ & \times \left(\mathsf{E}[V_{t+1} \mid H_t, \text{investor action at } t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1})] - \mathsf{bid}_{t+1} - \mathrm{fee}^{L}_{\mathrm{LLT}}\right),\end{aligned}$$

# 2 Equilibrium: No Trading Fees

In this section, we assume that traders (both, investors and low-latency liquidity providers) incur no trading fees.

## 2.1 Pricing and Decision Rules

**Equilibrium Pricing Rule.** We look for an equilibrium, in which low-latency liquidity providers post competitive limit orders and make zero profits, in expectation. We denote the equilibrium bid and ask prices at period $t$ by $\mathsf{bid}^*_t$ and $\mathsf{ask}^*_t$, respectively, and we

[9] With unit demands of investors, a low-latency trader has no incentive to post a limit order "into a queue": a market sell order that executes against the "first in the queue" order is informative, thus the liquidity provider will not want to modify her "second in the queue" order upon execution of the first.

use $\mathrm{MB}_t^*$ and $\mathrm{MS}_t^*$ denote, respective, the period $t$ investor's decisions to submit a market buy order price $\mathsf{ask}_t^*$ and a market sell order at price $\mathsf{bid}_t^*$.

The low-latency liquidity provider payoffs, given by equation (4), then implies the following competitive equilibrium pricing rules:

$$\mathsf{bid}_t^* = \mathsf{E}[V_t \mid H_t, \mathrm{MS}_t(\mathsf{bid}_t^*)] \tag{5}$$

$$\mathsf{ask}_t^* = \mathsf{E}[V_t \mid H_t, \mathrm{MB}_t(\mathsf{ask}_t^*)], \tag{6}$$

where we used the fact that history $H_{t-1}$ together with the period $t-1$ investor's action yield the same information about the security's value $V_t$ as history $H_t$ (because information about $V_t$ is only publicly revealed through investors' actions).

**Investor Actions with Competitive Liquidity Provision.** We focus on investor choices to buy; sell decisions are analogous. An investor can choose to submit a market order or a limit order, and, if he chooses to submit a limit order, technically, he may also choose the limit price. We search for an equilibrium where low-latency liquidity providers ensure that bid and ask prices are set competitively and equal the expected security value, conditional on the information available to the low-latency liquidity providers. An investor's choice of the limit price is thus mute, since a limit order that is posted at a price other than the prescribed, competitive equilibrium prices either yields the submitter negative profits in expectation or does not execute, because of the presence of low-latency traders. Because an investor is always able to obtain a zero profit by abstaining from trade, we restrict attention to limit orders posted at the competitive, equilibrium prices.

**Non-Competitive Limit Orders.** Formally, the zero probability of execution for limit orders posted at non-competitive prices is achieved by defining appropriate beliefs of market participants, regarding the information content of a limit order that is posted at an "out-of-the-equilibrium" price (e.g., when the period $t$ investor posts a limit order to buy at a price different from $\mathsf{bid}_{t+1}^*$) — so-called out-of-equilibrium beliefs. The appropriate definition of out-of-equilibrium beliefs is frequently necessary to formally describe equilibria with asymmetric information. To see the role of these beliefs in our model, observe first that when an order is posted at the prescribed, competitive equilibrium price, market participants derive the order's information content by Bayes' Rule, using their knowledge of equilibrium strategies. The knowledge of equilibrium strategies, however, does not help market participants to assess the information content of an order that cannot occur in equilibrium — instead, traders assess such an order's information content using out-of-the-equilibrium beliefs. We describe these beliefs in Appendix A, and we focus on prices and actions that occur in equilibrium in the main text.

**Investor Equilibrium Payoffs.** Because innovations to the fundamental are independent across periods, all market participants interpret the transaction history in the same manner. A period $t$ investor decision then does not reveal any additional information about innovations $\delta_\tau$, for $\tau < t$, and the equilibrium pricing conditions (5)-(6) can be written as

$$\begin{aligned} \mathsf{bid}_t^* &= \mathsf{E}[V_{t-1} \mid H_t] + \mathsf{E}[\delta_t \mid H_t, \mathrm{MS}_t(\mathsf{bid}_t^*)] && (7) \\ \mathsf{ask}_t^* &= \mathsf{E}[V_{t-1} \mid H_t] + \mathsf{E}[\delta_t \mid H_t, \mathrm{MB}_t(\mathsf{ask}_t^*)] && (8) \end{aligned}$$

The independence of innovations across time further allows us to decompose investors' expectations of the security's value, to better understand investor equilibrium payoffs. The period $t$ investor's expectation of the security's value at period $t$ is given by

$$\mathsf{E}[V_t \mid \delta_t, H_t] = \delta_t + \mathsf{E}[V_{t-1} \mid H_t]. \tag{9}$$

When the period $t$ investor submits a limit order to buy, his order will be executed at period $t+1$ (or never), and we thus need to understand this investor's expectation of the time $t+1$ value, conditional on his private and public information and on the order execution, $\mathsf{E}[V_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}^*)]$. Since the decision of the period $t+1$ investor reveals no additional information regarding past innovations, we thus obtain

$$\mathsf{E}[V_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}^*_{t+1})] = \mathsf{E}[V_{t-1} \mid H_t] + \delta_t + \mathsf{E}[\delta_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}^*_{t+1})] \quad (10)$$

Further, the independence of innovations implies that, conditional on the period $t$ investor submitting a limit buy order at price $\mathsf{bid}^*_{t+1}$, the period $t$ investor's private information of the innovation $\delta_t$ does not afford him an advantage in estimating the innovation $\delta_{t+1}$ or the probability of a market order to sell at period $t + 1$, relative to the information $H_{t+1}$ that will be publicly available at period $t+1$ (including the information that will be revealed by the period $t$ investor's order). Consequently, the period $t$ investor's expectation of the innovation $\delta_{t+1}$ coincides with the corresponding expectation of the low-latency liquidity providers, conditional on the period $t$ investor's limit buy order at price $\mathsf{bid}^*_{t+1}$.

The above insight, together with expressions (2)-(3) and (7)-(10), implies that an investor's expected payoffs to submitting market and limit buy orders, respectively, can be written as

$$\pi_t^{MB}(y_t, \delta_t) \; = \; y_t + \delta_t - \mathsf{E}[\delta_t \mid H_t, \mathrm{MB}_t(\mathsf{ask}^*_t)] \quad (11)$$

$$\pi_t^{LB}(y_t, \delta_t) \; = \; \mathsf{Pr}(\mathrm{MS}_{t+1}(\mathsf{bid}^*_{t+1}) \mid \mathrm{LB}_t(\mathsf{bid}^*_{t+1}), H_t) \left(y_t + \delta_t - \mathsf{E}[\delta_t \mid \mathrm{LB}_t(\mathsf{bid}^*_{t+1}), H_t]\right) (12)$$

**Investor Equilibrium Decision Rules.** An investor submits an order to buy if, conditional on his information *and* on the submission of his order, his expected profits are non-negative. Moreover, conditional on the decision to trade, an investor chooses the order type that maximizes his expected profits. An investor abstains from trading if he expects to make negative profits from all order types.

Expressions (11)-(12) illustrate that the period $t$ investor payoffs, conditional on the order execution, are determined by this investor's informational advantage with respect to the period $t$ innovation to the fundamental value (relative to the information content revealed by the investor's order submission decision) and by the investor's private valuation of the security. Our model is stationary, and in what follows, we restrict attention to investor decision rules that are independent of the history but are *solely* governed by an investor's private valuation and his knowledge of the innovation to the security's value.

When the decision rules at period $t$ are independent of the history $H_t$, the public expectation of the period $t$ innovation, conditional on the period $t$ investor's action, does not depend on the history either. Expressions (11)-(12) reveal that neither do investor equilibrium payoffs. Our setup is thus internally consistent in the sense that the assumed stationarity of the investor decision rules does not preclude investors from maximizing their payoffs.

Expected payoffs of a period $t$ investor are affected by the realizations of his private value $y_t$ and the innovation $\delta_t$ only through the sum of this investor's realized private value $y_t$ and his expectation of $\delta_t$, conditional on the period $t$ investor's information. We thus focus on decision rules with respect to this sum, and we refer to it as the *aggregate valuation*, and we denote the period $t$ investor's aggregate valuation by

$$z_t = y_t + \delta_t. \tag{13}$$

The aggregate valuation $z_t$ is symmetrically distributed on the interval $[-2, 2]$.

## 2.2 Equilibrium Characterization

We first derive properties of market and limit orders that must hold in equilibrium.

Our setup is symmetric, and we focus on decision rules that are symmetric around the zero aggregate valuation, $z_t = 0$. We focus on equilibria where investors use both limit and market orders.[10] Appendix A establishes the following result on the market's

[10] Any equilibrium where low-latency liquidity providers are the only liquidity providers closely resembles equilibria in market maker models in the tradition of Glosten and Milgrom (1985). In such an equilibrium, trading roles are pre-defined and maker-taker fees have no economic impact. We discuss further details in the Supplementary Appendix.

reaction to market and limit orders.

**Lemma 1 (Informativeness of Trades and Quotes)** *In an equilibrium where investors use both limit and market orders, both trades and investors' limit orders contain information about the security's fundamental value; a buy order increases the expectation of the security's value and a sell order decreases it.*

Lemma 1 implies that a price improvement stemming from a period $t$ investor's limit buy order at the equilibrium price $\mathsf{bid}^*_{t+1} > \mathsf{bid}^*_t$ increases the expectation of a security's value. In our setting, such a buy order will be immediately followed by a cancellation of a sell limit order at the best period $t$ price $\mathsf{ask}^*_t$ and a placement of a new sell limit order at the new ask price $\mathsf{ask}^*_{t+1} > \mathsf{ask}^*_t$ by a low-latency liquidity provider.

**Lemma 2 (Equilibrium Market and Limit Order Submission)** *In any equilibrium with symmetric time-invariant strategies, investors use threshold strategies: investors with the most extreme aggregate valuations submit market orders, investors with moderate aggregate valuations submit limit orders, and investors with aggregate valuations around 0 abstain from trading.*

To understand the intuition behind Lemma 2, observe first that, conditional on order execution, an investor's payoff is determined, loosely, by the advantage that his aggregate valuation provides relative to the information revealed by his order (see expressions (11)-(12)). Second, since market orders enjoy guaranteed execution, whereas limit orders do not, for limit orders to be submitted in equilibrium, the payoff to an executed limit order must exceed that of an executed market order. Consequently, the public expectation of the innovation $\delta_t$, conditional on, say, a limit buy order at period $t$, must be smaller than the corresponding expectation, conditional on a market buy order at period $t$ (in other words, the price impact of a limit buy order must be smaller than that of a market buy order). For this ranking of price impacts to occur, investors who submit limit orders must, on average, observe lower values of the innovation than investors who submit market buy orders. With symmetric distributions of both, the innovations and investor private values, we arrive at the previous lemma.

## 2.3 Equilibrium Existence

Utilizing Lemmas 1 and 2, we look for threshold values $z^M$ and $z^L < z^M$ such that investors with aggregate valuations above $z^M$ submit market buy orders, investors with aggregate valuations between $z^L$ and $z^M$ submit limit buy orders, investors with aggregate valuations between $-z^L$ and $z^L$ abstain from trading. Symmetric decisions are taken for orders to sell. Investors with aggregate valuations of $z^M$ and $z^L$ are marginal, in the sense that the investor with the valuation $z^M$ is indifferent between submitting a market buy order and a limit buy order, and the investor with the valuation $z^L$ is indifferent between submitting a limit buy order and abstaining from trading. Using (11)-(12), and the definition of the aggregate valuation (13), thresholds $z^M$ and $z^L$ must solve the following equilibrium conditions

$$\begin{aligned} z^M - \mathsf{E}[\delta_t \mid \mathrm{MB}_t] &= \mathsf{Pr}(\mathrm{MS}_{t+1}) \left(z^M - \mathsf{E}[\delta_t \mid \mathrm{LB}_t]\right) \qquad (14) \\ z^L &= \mathsf{E}[\delta_t \mid \mathrm{LB}_t], \qquad (15) \end{aligned}$$

where the stationarity assumption on investors' decision rules allows us to omit conditioning on the history $H_t$; $\mathrm{MB}_t$ denotes a market buy order at period $t$, which occurs when the period $t$ investor aggregate valuation $z_t$ is above $z^M$ ($z_t \in [z^M, 2]$), $\mathrm{LB}_t$ denotes a limit buy order at period $t$ ($z_t \in [z^L, z^M)$), and $\mathrm{MS}_{t+1}$ denotes a market order to sell at period $t+1$ ($z_{t+1} \in [-2, -z^M]$). Given thresholds $z^M$ and $z^L$, these expectations and probabilities are well-defined and can be written out explicitly, as functions of $z^M$ and $z^L$ (and independent of the period $t$).

Further, when investors use thresholds $z^M$ and $z^L$ to determine their decision rules, the bid and ask prices that yield zero profits to low-latency liquidity providers, given by the expressions in (7)-(8), can be expressed as

$$\begin{aligned} \mathsf{bid}_t^* &= p_{t-1} + \mathsf{E}[\delta_t \mid z_t \leq -z^M] \quad (16) \\ \mathsf{ask}_t^* &= p_{t-1} + \mathsf{E}[\delta_t \mid z_t \geq z^M], \quad (17) \end{aligned}$$

where $p_{t-1} \equiv \mathsf{E}[V_{t-1}|H_t]$. The choice of notation for the public expectation of the security's value recognizes that this expectation coincides with a transaction price in period $t-1$ (when such a transaction occurs). Expanding the above expressions one step further, for completeness, investors who submit limit orders to buy and sell at period $t$, in equilibrium, will post them at prices $\mathsf{bid}_{t+1}^*$ and $\mathsf{ask}_{t+1}^*$, respectively, given by

$$\begin{aligned} \mathsf{bid}_{t+1}^* &= p_{t-1} + \mathsf{E}[\delta_t \mid z_t \in [z^L, z^M)] + \mathsf{E}[\delta_{t+1} \mid z_{t+1} \leq -z^M] \quad (18) \\ \mathsf{ask}_{t+1}^* &= p_{t-1} + \mathsf{E}[\delta_t \mid z_t \in (-z^M, -z^L]] + \mathsf{E}[\delta_{t+1} \mid z_{t+1} \geq z^M] \quad (19) \end{aligned}$$

Finally, note that since the innovations are distributed symmetrically around 0, the public expectation of the period $t$ value of the security at the very beginning of period $t$, $\mathsf{E}[V_t|H_t]$, equals $p_{t-1}$. We prove the following existence theorem in Appendix A:[11]

**Theorem 1 (Equilibrium Characterization and Existence)** *There exist threshold values $z^M$ and $z^L$, with $0 < z^L < z^M < 2$, that solve indifference conditions (14)-(15). These threshold values constitute an equilibrium for any history $H_t$, given competitive equilibrium prices, $\mathsf{bid}_t^*$ and $\mathsf{ask}_t^*$ in (16)-(17), for the following trader decision rules. The investor who arrives at period t with aggregate valuation $z_t$*

- *places a market buy order if $z_t \geq z^M$,*
- *places a limit buy order at price $\mathsf{bid}_{t+1}^*$ if $z^L \leq z_t < z^M$,*
- *abstains from trading if $-z^L < z_t < z^L$.*

*Investors' sell decisions are symmetric to buy decisions.*

[11] Appendix A further provides the out-of-the-equilibrium beliefs that support the equilibrium prices and decision rules, described in Theorem 1.

# 3 Equilibrium With Trading Fees

Limit order books are maintained by exchanges that charge fees for executing orders. In what follows, we study the so-called maker-taker fee system, now common practice in equity markets worldwide. Under this system, exchanges charge different fees for trading with market and limit orders. For most of our discussion, we focus on the prevalent practice where the exchange only charges traders to remove, or take, liquidity and subsidizes traders who provide, or make, liquidity. The fee levied on market order submitters is referred as the "taker fee", and the rebate paid to submitters of executed limit orders is referred to as the "maker rebate". The intuition for our results extends for the reverse scenario where market order submitters receive a rebate and submitters of executed limit orders pay a positive fee.[12] Exchange fees are independent of whether the order is a buy order or a sell order.

We further assume that investors (who trade for informational and liquidity reasons) submit their orders via broker, whereas low-latency liquidity providers access the market directly. Brokers submit all traders' orders to the limit order book for execution, pay taker fees on market orders and receive maker rebates on executed limit orders. We assume that brokers act competitively and make zero profits on an average trade. We compare two settings. In the benchmark model, brokers pass the taker fees and maker rebates to the investors on a trade-by-trade basis. In the second, arguably more realistic setting, brokers charge investors a flat fee per trade. We assume that this fee is set to be the average fee incurred by a broker for executing an investor's order. Low-latency liquidity providers connect to the exchange directly in both settings, and they receive maker rebates on a trade-by-trade basis.[13]

We denote the taker fee by $f^{ta}$ and the maker fee by $f^{ma}$. The total fee charged by

[12] This "inverted" pricing is often referred to by industry participants as "taker-maker pricing", as it is utilized, for instance, by NASDAQ OMX BX.

[13] Since low-latency liquidity providers only submit limit orders, they do not incur taker fees. The assumption of connecting directly is thus equivalent to them connecting through brokers who charge differential fees to low-latency liquidity providers, relative to the rest of the investors.

the exchange for an executed trade is $f^{total} = f^{ta} + f^{ma}$. When discussing the intuition for our results, we will focus on $f^{ta} > 0$ and $f^{ma} < 0$ (a rebate).

## 3.1 Benchmark Model: Investors Pay Maker-Taker Fees

We first assume that irrespective of their identity, submitters of market orders pay the taker fee and submitters of executed limit orders receive the rebate.

**Equilibrium Pricing Rule:** Low-latency traders continue to ensure competitive pricing in the limit order book and continue to make zero profits in expectation. With a positive maker rebate, liquidity providers are willing to pay more than the expected value of the security when buying, and they are willing to accept less than the expected value when selling the security. The zero profit equilibrium bid and ask prices, given by expressions (7)-(8) in the absence of trading fees, become

$$\begin{aligned} \mathsf{bid}_t^* &= \mathsf{E}[V_{t-1} \mid H_t] + \mathsf{E}[\delta_t \mid H_t, \mathrm{MS}_t(\mathsf{bid}_t^*)] - f^{ma} \qquad (20) \\ \mathsf{ask}_t^* &= \mathsf{E}[V_{t-1} \mid H_t] + \mathsf{E}[\delta_t \mid H_t, \mathrm{MB}_t(\mathsf{ask}_t^*)] + f^{ma} \qquad (21) \end{aligned}$$

Ceteris paribus, a rebate to submitters of executed limit orders narrows the bid-ask spread, $\mathsf{ask}_t^* - \mathsf{bid}_t^*$, by twice the amount of the rebate. An investor's expected payoffs to market and limit orders, given by (2)-(3), can be written as

$$\begin{aligned} \pi_t^{MB}(y_t, \delta_t) &= y_t + \mathsf{E}[V_t \mid \delta_t, H_t] - \mathsf{ask}_t - f^{ta}, \qquad (22) \\ \pi_t^{LB}(y_t, \delta_t, \mathsf{bid}_{t+1}) &= \mathsf{Pr}(\mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}) \mid \delta_t, H_t) \qquad (23) \\ &\quad \times \left(y_t + \mathsf{E}[V_{t+1} \mid \delta_t, H_t, \mathrm{MS}_{t+1}(\mathsf{bid}_{t+1})] - \mathsf{bid}_{t+1} - f^{ma}\right). \end{aligned}$$

Using conditions (20)-(21) on the equilibrium bid and ask prices, expressions (22)-(23) can be rewritten analogously to (11)-(12), to reveal that the maker-taker fees only affect an investor's expected payoffs through the total exchange fee $f^{total} = f^{ta} + f^{ma}$:

$$\begin{aligned} \pi_t^{MB}(y_t, \delta_t) &= y_t + \delta_t - \mathsf{E}[\delta_t \mid H_t, \mathrm{MB}_t(\mathsf{ask}_t^*)] - f^{total} & (24) \\ \pi_t^{LB}(y_t, \text{info on } \delta_t) &= \mathsf{Pr}(\mathrm{MS}_{t+1}(\mathsf{bid}_{t+1}^*)) \mid \mathrm{LB}_t(\mathsf{bid}_{t+1}^*), H_t) & (25) \\ &\quad \times \left(y_t + \delta_t - \mathsf{E}[\delta_t \mid \mathrm{LB}_t(\mathsf{bid}_{t+1}^*), H_t]\right). \end{aligned}$$

Since an investor's payoffs do not depend on the split between the taker fee and the maker rebate (and low-latency liquidity providers make zero profits), in this setting this split has no economically meaningful impact (but it does affect the quoted bid-ask spread), consistent with Colliard and Foucault (2012).

**Proposition 1 (Independence of the Maker-Taker Split)** *Investors' equilibrium strategies and payoffs only depend on the total fee charged by the exchange, $f^{total} = f^{ta} + f^{ma}$.*

## 3.2 The Flat Fee Model

We now study the market where brokers do not pass through the taker fees and maker rebates, but instead charge investors a flat fee per trade, instead of passing through taker fees and maker rebates per trade. We assume that brokers act competitively, in the sense that they charge each investor the fee that yields the brokers zero profits in expectation. Since the limit order book is always full, the period $t$ investor's market order will incur a taker fee with certainty, and a period $t$ investor's limit order to buy (sell) will receive a maker rebate, provided that a market order to sell (buy) is submitted in period $t+1$. The expected fee $\bar{f}_t$ that the broker pays to the exchange for the period $t$ investor's order, conditional on the execution of this order, is then given by

$$\bar{f}_t = \frac{f^{ta} \cdot [\mathsf{Pr}(\mathrm{MB}_t^*) + \mathsf{Pr}(\mathrm{MS}_t^*)] + f^{ma} \cdot [\mathsf{Pr}(\mathrm{LB}_t^*) \cdot \mathsf{Pr}(\mathrm{MS}_{t+1}^*) + \mathsf{Pr}(\mathrm{LS}_t) \cdot \mathsf{Pr}(\mathrm{MB}_{t+1}^*)]}{\mathsf{Pr}(\mathrm{MB}_t^*) + \mathsf{Pr}(\mathrm{MS}_t^*) + \mathsf{Pr}(\mathrm{LB}_t^*) \cdot \mathsf{Pr}(\mathrm{MS}_{t+1}^*) + \mathsf{Pr}(\mathrm{LS}_t^*) \cdot \mathsf{Pr}(\mathrm{MB}_{t+1}^*)} \tag{26}$$

where $\mathrm{LB}_t^*$ and $\mathrm{MB}_t^*$ denote the period $t$ investor's market and limit orders to buy at the equilibrium bid and ask prices; likewise for the sell orders and orders in period $t+1$.

As in Section 2, we focus on an equilibrium where investors use stationary, time-invariant threshold strategies with respect to their aggregate valuation $z_t = y_t + \delta_t$. Because innovations $\delta_t$ to the security's value and investor private valuations $y_t$ are identically and independently distributed across time, probabilities of market and limit orders to buy and to sell are time-invariant. We continue to focus on a symmetric equilibrium, where investors decisions to buy and sell are symmetric with respect to the aggregate valuation $z_t = 0$, so that the probability of a market buy order then equals the probability of a market sell order; likewise for limit orders. Consequently, the expected per-investor fee does not depend on period $t$. Denoting this fee by $\bar{f}$ and writing $\mathsf{Pr}(\mathrm{LB}^*)$ for the probability of a limit (buy) order in equilibrium, we simplify (26) to

$$\bar{f} = \frac{f^{ta} + f^{ma} \cdot \mathsf{Pr}(\mathrm{LB}^*)}{1 + \mathsf{Pr}(\mathrm{LB}^*)} \tag{27}$$

Since low-latency liquidity providers receive maker rebates and act competitively, limit order book prices are determined by the same conditions as in the benchmark model (conditions (20)-21)). Investor payoffs, however, are affected by the flat fee $\bar{f}$. With the decision rules being stationary, these payoffs are given by

$$\begin{aligned}
\pi^{\mathrm{MB}}(y_t, \delta_t) &= y_t + \delta_t - (\mathsf{E}[\delta_t \mid \mathrm{MB}_t^*] + f^{ma}) - \bar{f} && (28)\\
\pi^{\mathrm{LB}}(y_t, \delta_t) &= \mathsf{Pr}(\mathrm{MS}_{t+1}^* \mid \mathrm{LB}_t^*) \left( y_t + \delta_t - (\mathsf{E}[\delta_t \mid \mathrm{LB}_t^*] - f^{ma}) - \bar{f} \right), && (29)
\end{aligned}$$

where $\mathrm{LB}_t^*$ and $\mathrm{MB}_t^*$ denote investors' limit and market buy orders at the equilibrium competitive prices; the stationarity of investor decision rules allows us to drop the dependence on the history. Substituting in the expression for the zero-profit flat fee charged by brokers and using $f^{ta} + f^{ma} = f^{total}$, we obtain

$$
\begin{aligned}
\pi^{\mathrm{MB}^*}(y_t, \delta_t) &= y_t + \delta_t - \mathsf{E}[\delta_t \mid \mathrm{MB}_t^*] - \frac{f^{total} + 2f^{ma} \cdot \mathsf{Pr}(\mathrm{LB}_t^*)}{1 + \mathsf{Pr}(\mathrm{LB}_t^*)} && (30) \\
\pi^{\mathrm{LB}^*}(y_t, \delta_t) &= \mathsf{Pr}(\mathrm{MS}_{t+1} \mid \mathrm{LB}_t^*) && (31) \\
&\quad \times \left( y_t + \delta_t - \mathsf{E}[\delta_t \mid \mathrm{LB}_t^*] - \frac{f^{total} - 2f^{ma}}{1 + \mathsf{Pr}(\mathrm{LB}_t^*)} \right).
\end{aligned}
$$

Equations (30)-(31) illustrate, in particular, that when only investors pay a flat fee per trade, their payoffs are affected by the maker (or taker) fee beyond the effect of the total exchange fee. The split between the taker fee and the maker rebate will thus be economically relevant in this setting.

## 3.3 Equilibrium Characterization

Colliard and Foucault (2012) analyze the impact of the total fee on trader behavior. We focus instead on the split of the exchange fee into the taker fee and the maker rebate. In what follows, we set the total fee that the exchange charges to 0, so that $f^{ta} = -f^{ma}$, and we use the notation $f \equiv f^{ta}$.

**The Benchmark Model.** When the total fee charged by the exchange is set to 0, the benchmark model is economically equivalent to the model in absence of fees, described in Section 2, in the sense that conditions that the defined threshold decision rules for investors are the same. To see this, observe that when $f^{total} = 0$, equations (24)-(25) that determine investor payoffs are the same as equations (11)-(12). Equilibrium conditions on investor thresholds in the benchmark model then coincide with those in the absence of fees (conditions (14)-(15)).

The only difference between the no-fee setting to the benchmark model is the quoted bid-ask spread. Specifically, for any prior expectation $p_{t-1} = \mathsf{E}[V_t|H_t]$ of the security's value $V_t$, the bid and ask equilibrium prices in the benchmark model with fees satisfy

$$\begin{aligned} \mathsf{bid}_t^* &= p_{t-1} + \mathsf{E}[\delta_t \mid z_t \leq -z^M] + f & (32) \\ \mathsf{ask}_t^* &= p_{t-1} + \mathsf{E}[\delta_t \mid z_t \geq z^M] - f, & (33) \end{aligned}$$

where, as before, we use $z^M$ to denote the threshold aggregate valuation of the investor who is indifferent between submitting a limit order and a market order (investors with valuations $z_t$ above $z^M$ submit market buy orders, investors with $z_t$ below $-z^M$ submit market sell orders). When $f = 0$, conditions (32)-(33) coincide with conditions (16)-(17) on the equilibrium bid and ask prices in the absence of the fees.

In the absence of fees, the bid-ask spread is positive as long as market orders are informative. When $f \neq 0$, however, this is no longer the case. Equations (32)-(33) imply that in a symmetric equilibrium, $\mathsf{ask}_t^* - \mathsf{bid}_t^* > 0$ if and only if

$$f < \mathsf{E}[\delta_t \mid z_t \geq z^M]. \tag{34}$$

**Proposition 2 (Existence in the Benchmark Model)** *There exist values $z^M$ and $z^L$, with $0 < z^L < z^M < 2$, that solve indifference conditions (14)-(15). These values together with equilibrium prices $\mathsf{bid}_t^*$ and $\mathsf{ask}_t^*$ given by (32)-(33) constitute an equilibrium, with decision rules described in Theorem 1, if and only if condition (34) is satisfied.*

**The Flat Fee Model.** With $f = f^{ta} = -f^{ma}$, the flat fee (expression (27)) is

$$\bar{f} = \frac{1 - \mathsf{Pr}(\mathrm{LB}^*)}{1 + \mathsf{Pr}(\mathrm{LB}^*)} \cdot f \tag{35}$$

Expression (35) illustrates, in particular, that the flat fee has the opposite sign of the maker rebate. In particular, when the maker rebate is positive, brokers always set a positive flat fee (despite the zero total fee). The presence of low-latency liquidity providers ensures that market orders always execute, whereas limit orders only execute when another investor submits a market order. Low-latency liquidity providers must capture a fraction of the maker rebates, leaving investors to pay a positive exchange fee.

**Lemma 3 (Flat Fee)** *The flat fee $\bar{f}$ set by brokers is positive when the maker rebate is positive, and it is negative when the maker rebate is negative.*

Our further results on the flat fee model are numerical. We employ the following family of distributions of the innovation parameter $\delta_t$, for $\alpha \geq 1$.[14]

$$\bar{g}(\delta, \alpha) = \begin{cases} \frac{(1-\delta)^{(\alpha-1)}}{\alpha} & \text{if } \delta \geq 0 \\ \frac{(1+\delta)^{(\alpha-1)}}{\alpha} & \text{if } \delta \leq 0 \end{cases} \tag{36}$$

The distribution family includes the uniform distribution ($\alpha = 1$).

We numerically search for an equilibrium, with properties similar to those in Section 2. Specifically, we look for an equilibrium where investors use threshold rules that are symmetric and that do not depend on the history, such that investors with most extreme aggregate valuations trade with market orders, investors with moderate aggregate valuations trade with limit orders, and investors with aggregate valuations around 0 abstain from trading. The equilibrium indifference conditions are analogous to conditions (14)-(15), except that they are adjusted for the exchange fees, using (30)-(31):[15]

$$\begin{aligned} z^M - \mathsf{E}[\delta_t \mid \mathrm{MB}_t] + f - \bar{f} &= \mathsf{Pr}(\mathrm{MS}_{t+1}) \left( z^M - \mathsf{E}[\delta_t \mid \mathrm{LB}_t] - f - \bar{f} \right) \\ z^L &= \mathsf{E}[\delta_t \mid \mathrm{LB}_t] + f + \bar{f}. \end{aligned} \tag{37}$$

# 4 Impact of Fees on Liquidity and Volume

We continue to assume that the total fee charged by the exchange is set to 0, so that the maker rebate equals the taker fee, $f^{ta} = -f^{ma} = f$. We analyze the impact of an increase in the maker rebate (and the taker fee), measured by an increase in $f$, on

[14] Density $2\bar{g}$ is a Beta-distribution on [0,1].

[15] Numerically, the solution is always unique. If it were not unique, we would focus on the one that delivers the smallest bid-ask spread in equilibrium.

quoted and cum-fee bid-ask spreads, trading volume, and market participation. The quoted bid-ask spread is the difference between the ask and bid prices. The cum-fee spread additionally accounts for the fee paid by a submitter of a market order; this fee is the taker fee in the benchmark model and the flat fee $\bar{f}$ in the flat fee model. We measure market participation by the probability that an investor does not abstain from submitting an order, and we measure trading volume by the probability that an investor submits a market order (since market orders always execute in our setting).

Proposition 2 implies the following result for the benchmark model.

**Corollary 1 (Impact of Fees in the Benchmark Model)** *In an equilibrium of the benchmark model, thresholds $z^M$ and $z^L$, market participation, trading volume, and cum-fee bid-ask spreads are independent of $f$. Quoted bid ask-spreads decline in $f$.*

The values of observable variables in the benchmark model coincide with the corresponding values in the flat fee model for $f = 0$. The discussion below examines the flat fee model for the case of a zero total exchange fee.

**Trading Volume and Market Participation.** Equations (28)-(29), which define investor payoffs in a flat fee model, illustrate that, ceteris paribus, an increase in the maker rebate provides investors with incentives to switch from limit to market orders. All else equal, such an increase will decrease the spread, thus increasing the payoff to market orders and simultaneously reducing the payoff to limit orders. In contrast to the benchmark model, however, changes in the bid-ask spread are not offset by the changes in investor fees — because the flat fee charged by brokers does not depend on the order type. Since trade occurs in our model when a market order is submitted, an increase in the probability of a market order implies an increase in trading volume.

The impact on investors who were previously indifferent between submitting a limit order and abstaining from trading is more complex. On the one hand, ceteris paribus, as traders increase their usage of market orders, limit orders are submitted by less informed traders, the price impact of a limit order declines, and limit orders become more attractive. On the other hand, an increase in the maker rebate leads to a decline in the bid-ask spread, making limit order prices less attractive to investors who do not receive the rebate. Numerical simulations reveal that the latter effect dominates in our setting; that is, market participation declines.

**What happens when the maker rebate is very large?** As the taker fee and the maker rebate increase, threshold $z^M$ decreases and threshold $z^L$ increases. When the maker rebate is sufficiently high (relative to the spread), a limit order yields negative profits to investors in expectation, because they do not receive maker rebates. When this happens, low-latency liquidity providers become the only submitters of limit orders, while investors trade exclusively with market orders. As a consequence, the flat fee equals the taker fee. The marginal submitter of a market order is then exactly indifferent between submitting a market order and abstaining from trading, and he earns zero expected profits. We denote the aggregate valuation of such a marginal submitter by $z_0$, and the value of $f$ that yields $z^M = z^L = z_0$ in equilibrium by $f_0$. Using investor payoffs, given by expression (28), together with $\bar{f} = f^{ta} = -f^{ma}$, we find that $z_0$ solves

$$z_0 - \mathsf{E}[\delta_t | z_t \geq z_0] = 0. \tag{38}$$

A further increase in the maker rebate (above $f_0$) then leads to a further decline in the quoted spread but does not have an effect on investors payoffs, because a decline in the quoted spread is exactly offset by an increase in the average fee, which equals the taker fee. As with the benchmark model, an equilibrium fails to exist when the maker rebate is so large that the bid-ask spread becomes nonpositive. Similarly to condition (34) for the benchmark model, the bid-ask spread remains positive for fees $f$ that are below value $f_1$ that solves

$$f_1 = \mathsf{E}[\delta_t \mid z_t \geq z_0^M] \tag{39}$$

**What happens when the maker rebate is negative?** When $f = f^{ta} = -f^{ma} < 0$, i.e. limit order submitters pay a positive fee for executed orders, whereas market order submitters receive a positive taker rebate, liquidity providers offer less than the expected value of the security when buying and they demand more than the expected value when selling the security. As a consequence, as the maker fee $(-f)$ increases from 0, quoted spreads widen. Investors pay a flat fee (in this case, the fee is negative, so they receive a flat positive rebate), therefore market orders become less attractive to them and limit orders become more attractive.

Intuitively, when the maker fee is positive and high ($f$ is low and negative), the bid ask spread becomes too wide, market orders earn negative profits for all investors (even after accounting for the positive flat rebate that investors receive on each transaction), and trade does not occur. We denote the level of the taker fee $f$ where this occurs by $f^{NT}$.

Furthermore, as the positive maker fee $(-f > 0)$ increases from 0 and the spread widens, investors who receive a positive flat rebate per transaction, irrespective of the order type, are more willing to submit limit orders. In particular, there may exist values of $f < 0$ such that an investor with the aggregate valuation of 0 is willing to submit a limit order — because the bid price that he would pay in the event his buy limit order executes is lower than indicated by the average information content of limit orders (because of the positive maker fee imposed on the low-latency liquidity providers) and additionally this investor receives a positive flat rebate. Our numerical simulations show that $f^{NT}$ is below this value, i.e., that there is a range of values of the taker fee such that all investors participate in the market (by trading with a limit or a market order).

Figure 2 illustrates the following observation on order submission decisions.

**Numerical Observation 1 (Fee Thresholds and Investor Equilibrium Actions)** *There exist $f^{NT}$, $f_0$, $f_1$, with $f^{NT} < 0 < f_0 < f_1$, such that in the flat fee model*

*(i) investors submit both market and limit orders in equilibrium with $f < f_0$;*

*(ii) investors submit only market orders in equilibrium when $f_0 \leq f < f_1$;*

*(iii) a stationary equilibrium with trade does not exist when $f \geq f_1$ or $f < f^{NT}$.*

*Threshold $f^{NT}$ is the highest value of $f$ that yields $z^M = 2$, $f_0$ is the value of $f$ that yields solutions $z^M = z^L = z_0^M$ to equations (37), and threshold $f_1$ solves (39).*

Figure 3 illustrates the following observation on probabilities of order submissions and the implications for trading volume and market participation.

**Numerical Observation 2 (Volume and Market Participation)** *In the flat fee model, as the taker fee increases and the maker decreases ($f$ increases), for $f^{NT} < f \leq f_0$, the probability that an investor*

*(i) submits a market order increases (trading volume increases);*

*(ii) submits a limit order decreases;*

*(iii) abstains from trading (weakly) increases (market participation declines).*

*These probabilities do not depend on $f$ when $f_0 \leq f < f_1$.*

**Quoted Bid-Ask Spread.** As the maker rebate increases ($f$ increases), more investors submit market orders, that is they submit aggressive orders for lower values of the innovations $\delta_t$. Furthermore, as $f$ increases, the bid-ask spread declines because low-latency liquidity providers compete the benefits of the increased rebate away. Both of these effects lead to a decline in the quoted bid-ask spread.

**Cum-Fee Bid-Ask Spread.** The cum-fee spread accounts for the fee that an investor pays to his broker:

$$\textit{cum-fee spread} = \mathsf{ask}_t^* - \mathsf{bid}_t^* + 2\bar{f}, \tag{40}$$

where the factor 2 accounts for the fact that the bid-ask spread is a cost of a round-trip transaction, so that the fee is paid twice. As the maker rebate increases ($f$ increases), the probability of a limit order declines, and expression (35) reveals that $\bar{f}$ increases as long as $f < f_0$. Numerically, this increase is more than offset by the decline in the quoted spread, so that the cum-fee spread declines. Figure 4 illustrates the following observation

**Numerical Observation 3 (Quoted and Cum-Fee Spreads: Flat Fee)** *As the taker fee and the maker rebate increase ($f$ increases), for $f^{NT} < f < f_1$,*

*(i) the quoted bid-ask spread declines;*

*(ii) the broker flat fee $\bar{f}$ increases;*

*(iii) the cum-fee spread declines for $f < f_0$ and is independent of $f$ for $f \geq f_0$.*

**Price Impact.** The price impact of a trade measures the change in the public expectation following the execution of a trade. In our model, this change is determined, loosely, by the information content of market orders about the time-$t$ innovation $\delta_t$. Specifically, the price impact of a buyer-initiated transaction is given by:

$$\text{price impact}_{\text{buy},t} = \mathsf{E}[V_t \mid \text{MB}_t] - p_{t-1} = \mathsf{E}[\delta_t \mid \text{MB}_t]. \tag{41}$$

Using expression (33) for the equilibrium ask price $\mathsf{ask}_t^*$, we find that for a positive maker rebate ( $f > 0$) the price impact of a trade is higher than indicated by a transaction price:

$$\text{price impact}_{\text{buy},t} = \mathsf{E}[\delta_t \mid \text{MB}_t] = \mathsf{ask}_t^* + f - p_{t-1} > \mathsf{ask}_t^* - p_{t-1}. \tag{42}$$

Figure 5 illustrates the above relation between the quoted half-spread, $\mathsf{ask}_t^* - p_{t-1}$.

Numerical Observation 2 illustrated, in particular, that as the taker fee $f$ increases from 0 to $f_0$, the marginal submitter of a market order requires a lower aggregate valuation. Consequently, market orders are submitted for lower absolute values of realizations of the time $t$ innovations $\delta_t$. This insight explains the following numerical observation, illustrated by Figure 5.

**Numerical Observation 4 (Price Impact: Flat Fee)** *The price impact of a trade is decreasing in $f$ on $[f^{NT}, f_0]$, and constant on $(f_0, f_1]$.*

Numerical Observation 4 is supported empirically by Malinova and Park (2011).

# 5 Impact of Fees on Welfare

When investors pay a flat fee per trade, an increase in the maker rebate $f$ reduces the zero-profit bid-ask spread. The decline in the bid-ask spread makes market orders relatively more attractive and limit orders relatively less attractive to investors who pay a flat fee per trade. As a result, some investors who would submit limit orders in the absence of maker-taker pricing choose to abstain from trading when the maker rebate is positive. At the same time, submitters of limit orders with higher valuations choose to switch to market orders, increasing the probability of execution for their orders, or the fill rate, (to certainty) and also for other investors' limit orders. While investors who choose to abstain from trading fail to realize their gains from trade, the remainder of limit order submitters realize gains from trade more frequently. The impact of maker-taker pricing on the aggregate welfare intuitively depends on whether welfare loss from a decline in market participation (by limit order submitters) is offset by a welfare gain stemming from an increase in the fill rate for the remainder of investors.

Each investor in our setting has a private valuation for the security, and we follow Bessembinder, Hao, and Lemmon (2012) to define a social welfare measure that reflects allocative efficiency. Specifically, we define welfare as the expected gain from trade in the market for a given period $t$. If a transaction occurs in period $t$, then the welfare gain is given by the private valuation of a buyer, net of the trading fee paid by the buyer, minus the private valuation of a seller, net of the trading fee paid by the seller.

A transaction in period $t$ occurs when the period $t$ investor submits a market buy or a market sell order. Focusing on a submitter of a buy market order: this investor trades with the period $t-1$ investor if the period $t-1$ investor submitted a limit sell order and he trades with a low-latency liquidity provider otherwise. With a flat fee set to equal the average fee paid by an investor, the expected aggregate fee on each transaction is zero. Accounting for the fact that a low-latency liquidity supplier has a zero private valuation, by symmetry, we obtain the following expression for the welfare (further details and explicit expressions are in the Appendix A):

$$W_t = 2 \cdot \mathsf{Pr}(MB_t) \left( \mathsf{E}[y_t \mid H_t, \mathrm{MB}_t] - \mathsf{Pr}(\mathrm{LS}_{t-1}) \cdot \mathsf{E}[y_{t-1} \mid H_{t-1}, \mathrm{MB}_t, \mathrm{LS}_{t-1}] \right). \qquad (43)$$

Proposition 2 implies the following result for the benchmark model.

**Corollary 2 (Welfare: Benchmark Model)** *In an equilibrium of the benchmark model, welfare is independent of $f$.*

Figure 6 illustrates the following observation on the impact of maker-taker pricing on social welfare.

**Numerical Observation 5 (Social Welfare: Flat Fee)** *Expected total welfare $W_t$ is increasing in $f$ on $[f^{NT}, f_0]$, and constant on $(f_0, f_1]$.*

In a world where exchange maker-taker fees are only passed through on average, our results suggest that positive maker rebates have a positive effect on social welfare. Allocative efficiency is highest when investors only trade with market orders (or not at all). An implication of our result on social welfare is that it is socially beneficial for investors and low-latency liquidity providers to specialize: investors submitting market orders, and; low-latency liquidity providers providing liquidity.

# 6 Conclusion

We develop a model to analyze a financial market where investors trade for informational and liquidity reasons in a limit order book that is permanently monitored by low-latency liquidity providers. We employ our model to study the impact of maker-taker fees, focussing on the current practice of the implementation of these fees. Maker-taker pricing, in its most common form, refers to a pricing scheme where exchanges pay traders a maker rebate to post liquidity and charge traders a positive taker fee to remove liquidity, and more, generally to a fee system that levies different fees for liquidity provision and removal.

We find that when all traders pay the maker-taker fees, investor behavior is affected only through the total fee charged by the exchange (the taker fee minus the maker rebate), consistent with Colliard and Foucault (2012). When, however, investors only pay the average maker-taker fee, through a "flat fee" per trade, the split of the total exchange fee into the maker fee and the taker fee also plays a meaningful role, because it differentially affects the incentives of low-latency liquidity providers and of investors. When the maker fee declines, low-latency liquidity providers quote lower bid-ask spreads. Consequently, investors who pay a flat fee per trade have an incentive to switch from limit orders to market orders.

The empirical predictions of our model support the industry's opinions on the impact of maker-taker pricing on long-term investors. Indeed, we predict that if a positive maker rebate is introduced (financed by an increase in the taker fee), investors trade on the liquidity demanding side more frequently, that they submit fewer limit orders, and that more of them choose to abstain from trading altogether. Our model also predicts an increase in the average exchange fee that a broker incurs when executing client orders, consistent with industry concerns. Contrary to industry opinions, we find that trading costs for liquidity demanders decrease, because a decline in the quoted spreads more than offsets the increase in the average exchange fee. One key contributor to the decline in trading costs for liquidity demanders is the decrease in price impact of trades — they become less informative, as less-informed investors trade aggressively, using market orders. Malinova and Park (2011) find empirical support for our predictions.

When the exchange charges a positive maker rebate, but brokers charge an average flat fee, maker-taker pricing affects investors' order choices and thus allocative efficiency. We find that an increase in the maker rebate leads investors to realize gains from trade more frequently, resulting in a positive maker rebate being socially optimal.

Our results have several policy implications. First, we find that in markets where brokers charge investors a flat fee per trade, the levels of maker and taker fees has an economic effect beyond that of the total exchange fee. A decrease in the maker fee decreases trading costs for market order submitters. When the fee is negative and sufficiently low, an equilibrium fails to exist in our model, as the bid-ask spread declines to zero. Our predictions may thus shed light on locked markets, where a bid price in one market equals the ask price in another. Our results suggest that locked markets occur more frequently when the maker rebates/taker fees are sufficiently large and that locked markets may arise, for instance, when low-latency liquidity providers post only bid quotes in one market and only ask quotes in the other.

Second, our results show that competition among brokers is not sufficient to neutralize the impact of the maker-taker fees — when the fee is passed through on average, investors' trading incentives are different to the situation where investors pay taker fees and receive maker rebates for each executed trade.

Third, we reiterate the prevailing academic opinion on the importance of accounting for the exchange trading fees (See, e.g., Angel, Harris, and Spatt (2011), Colliard and Foucault (2012), or Battalio, Shkilko, and Van Ness (2012).) A lower quoted spread need not imply lower trading costs for investors, and, consequently routing orders to the trading venue that is quoting the best price need not guarantee the best execution.

Fourth, we caution that the causal relations among trading volume, trading costs, and competition for liquidity providers are more complex than the taken-at-face-value intuition would suggest. An increase in volume in our setting is driven by changes in investor trading behavior. These changes necessitate a higher rate of participation by low-latency liquidity providers, which may manifest empirically as an increase in competition among low-latency liquidity providers.[16] Hence, an empirically observed increase in competition need not be the driving force of changes in trading volume

[16] In our model, low-latency liquidity providers compete in prices; empirical assessments typically measure competition in quantities.

and trading costs. Our results further highlight that trading volume in a limit order market, where some traders specialize in liquidity provision, is not determined by market participation of investors.

Our work focusses on the impact of maker-taker fees on investor trading behavior, and through it, on trading costs, market participation, volume, and social welfare. We acknowledge that several tradeoffs permit us to tractably analyze this impact, and our results must be interpreted with these tradeoffs in mind.

First, the tractability of our setup stems from the presence of low-latency liquidity providers: competition among them induces all limit order submitters to offer competitive prices and thus pins down limit order prices in equilibrium. Analyzing the impact of low-latency trader behavior on the remainder of the market participants is outside the scope of our model. Instead, our goal is to study a limit order market where low-latency traders are present and where their presence ensures competitive liquidity provision.

Second, we focus on investor trading incentives, assuming that brokers act competitively, and we study a single market. When markets are fragmented, brokers have a choice of where to send their client orders. Since trading fees differ across trading venues, a broker that charges investors a flat fee per trade, may have a conflict of interest with respect to the best execution for the client versus the lowest exchange fee. Such conflicts do not arise in our model, and we may thus understate investor trading costs.

# A Appendix

This Appendix provides proofs and necessary derivations that are omitted from the main part of the paper. It is preliminary, and it is incomplete in the current version of the paper. This version of the Appendix only provides a proof sketch for the existence theorem (Theorem 1); the intuition for the remainder of the results is in the main text.

## A.1 Preliminary Notation

Innovation $\delta_t$ is distributed on [-1,1], symmetrically around 0, according to the density function $\bar{g}$. On $[0,1]$, we have $\bar{g}(\cdot) = g(\cdot)/2$, where $g$ is a density function, and $g$ is declining. Denote the relevant distribution function by $G$. Since $g$ is declining, we obtain the following bounds on the density:

$$g(\delta) < \frac{G(\delta)}{\delta} \text{ and } g(\delta) > \frac{1 - G(\delta)}{1 - \delta}. \tag{44}$$

As in the main text, we denote the prior on the asset value at time $t$ by $v_t$, and we use $z_t$ to denote the period $t$ investor's aggregate valuation, $z_t = y_t + \delta_t$.

We will employ the following notation (spelled out for buys, sells are similarly), abusing it and omitting $t$ subscripts, since we are looking for a stationary equilibrium:

- For the expected innovation $\delta_t$, conditional on a market buy order at time $t$:

$$\mathsf{E}^M\delta := \mathsf{E}[\delta_t|\text{market buy at time } t] \tag{45}$$

- For the expected innovation $\delta_t$, conditional on a limit buy order at time $t$, which is posted at the competitive equilibrium price:

$$\mathsf{E}^L\delta := \mathsf{E}[\delta_t|\text{limit buy at time } t] \tag{46}$$

- For the probability of a market sell at time $t+1$:

$$\mathsf{pr}^M := \mathsf{Pr}[\text{market sell at } t+1] = \mathsf{Pr}[\text{market buy at } t+1] = \mathsf{Pr}[\text{market buy at } t] \tag{47}$$

- For the probability of a limit buy order at time $t$:

$$\mathsf{pr}^L := \mathsf{Pr}[\text{limit buy at } t] = \mathsf{Pr}[\text{limit sell at } t] \tag{48}$$

In what follows, we will treat $\mathsf{E}^L\delta$ and $\mathsf{E}^M\delta$ as *functions* of the relevant thresholds (as opposed to their equilibrium values).

## A.2 Proof of Theorem 1

Equilibrium thresholds solve equations (14). Using notation defined Section in A.1, the symmetry and the stationarity of the equilibrium that we are looking for, these conditions can be rewritten as

$$\begin{aligned} z^M - \mathsf{E}^M\delta &= \mathsf{pr}^M(z^M - \mathsf{E}^L\delta), \qquad (49) \\ z^L &- \mathsf{E}^L\delta. \qquad (50) \end{aligned}$$

An informed trader will submit a market buy over a limit buy as long as $z_t \geq z^M$, will submit a limit buy if $z^M > z_t \geq z^L$, and will abstain from trading otherwise. To show existence of a threshold equilibrium, we need to show existence of thresholds $z^M$ and $z^L$ and prove the optimality of trader strategies.

We proceed in 4 steps. In step 1, we show that for any given $z^M \in [0, {}^3\!/_4]$ there exists the unique $z^L$ that solves (50).[17] We denote this solution by $z^L_*(z^M)$ and show, in Step 2, that $z^L_*(z^M)$ is increasing in $z^M$. In Step 3, we show that there exists $z^M$ that solves

$$z^M - \mathsf{E}^M\delta = \mathsf{pr}^M(z^M - z^L_*(z^M)). \qquad (51)$$

Finally, in Step 4, we show the optimality of the strategies and discuss out-of-equilibrium beliefs that support these strategies in a perfect Bayesian equilibrium.

### A.2.1 Step 1: Existence and Uniqueness of $z^L_*(z^M)$

We first derive the expression for $\mathsf{E}^L\delta$ in terms of the model primitives:

$$\mathsf{E}^L\delta = \frac{\int_{-1}^{1} d\delta \int_{-1}^{1} dy(\delta \cdot h^L(\delta, y|\text{LB}))}{\int_{-1}^{1} d\delta \int_{-1}^{1} dy(h^L(\delta, y|\text{LB}))}, \qquad (52)$$

where function $h^L(\delta, y|\text{LB})$ is defined as follows:

$$h^L(\delta, y|\text{LB}) = \begin{cases} \frac{1}{2} \cdot \bar{g}(\delta), \text{ if } \delta \in [z^L - 1, 1] \text{ and } y \in [z^L - \delta, z^M - \delta] \\ 0, \text{ otherwise.} \end{cases} \qquad (53)$$

[17] Threshold ${}^3\!/_4$ may seem arbitrary, but we can also show that there does not exist $z^M > {}^3\!/_4$ that solves (49) for $\mathsf{E}^L\delta \geq 0$ )(i.e., when investors use both market and limit orders).

The denominator of (52) equals the probability of a limit buy order submission $\mathsf{pr}^L$, and we will use one more piece of short-hand notation:

$$\mathsf{num}(\mathsf{E}^L\delta) := \int_{-1}^{1} d\delta \int_{-1}^{1} dy(\delta \cdot h^L(\delta, y|\mathrm{LB})) \tag{54}$$

Using this notation, we then have $\mathsf{E}^L\delta = \mathsf{num}(\mathsf{E}^L\delta)/\mathsf{pr}^L$. Substituting $h^I$ in, putting in appropriate integral bounds and expressing $f$ as a function of $g$, we express the probability of a limit buy as follows:

$$\begin{aligned} \mathsf{pr}^L &= \frac{1}{4}\big(1 + G(1 - z^L)\big) \cdot (z^M - z^L) - \frac{1}{4}\int_{1-z^M}^{1-z^L} (\delta - (1 - z^M))g(\delta)d\delta \\ &\equiv \gamma^L \cdot (z^M - z^L) - \frac{1}{4}\int_{1-z^M}^{1-z^L} (\delta - (1 - z^M))g(\delta)d\delta, \end{aligned} \tag{55}$$

where $\gamma^L$ is defined accordingly. Note that, using this notation,

$$\frac{\partial \mathsf{pr}^L}{\partial z^L} = -\gamma^L. \tag{56}$$

Probability $\mathsf{pr}^L$ can also be expressed as

$$\mathsf{pr}^L = \gamma^M \cdot (z^M - z^L) + \frac{1}{4}\int_{1-z^M}^{1-z^L} (1 - z^L - \delta)g(\delta)d\delta, \tag{57}$$

where $\gamma^M \equiv {}^1\!/\!_4\mu(1 + G(1 - z^M))$. We then have

$$\frac{\partial \mathsf{pr}^L}{\partial z^L} = \gamma^M. \tag{58}$$

The numerator of the $\mathsf{E}^L\delta$ function can be expressed as

$$\mathsf{num}(\mathsf{E}^L\delta) = -\frac{1}{4}\int_{1-z^M}^{1-z^L} \delta(1 - z^L - \delta)g(\delta)d\delta + \frac{1}{4}(z^M - z^L)\int_{1-z^M}^{1} \delta g(\delta)d\delta, \tag{59}$$

where we used the following identity $z^M - z^L = (1 - z^L - \delta) + (z^M - 1 + \delta)$. Note that

$$\frac{\partial \mathsf{num}(\mathsf{E}^L\delta)}{\partial z^L} = -\frac{1}{4}\int\limits_{1-z^L}^{1} \delta g(\delta) d\delta \equiv -\beta^L, \text{ and} \tag{60}$$

$$\frac{\partial \mathsf{num}(\mathsf{E}^L\delta)}{\partial z^M} = \frac{1}{4}\int\limits_{1-z^M}^{1} \delta g(\delta) d\delta \equiv \beta^M, \text{ and} \tag{61}$$

**Lemma 4 (Bounds on $\mathsf{E}^L\delta$)** *Expectation $\mathsf{E}^L\delta$ satisfies $\mathsf{E}^L\delta < z^M/2$.*

*Proof Sketch:* We use the following bounds on expressions for $\mathsf{num}(\mathsf{E}^L\delta)$ and $\mathsf{pr}^L$:

$$\mathsf{num}(\mathsf{E}^L\delta) \leq \frac{1}{4}(z^M - z^L)\int\limits_{1-z^M}^{1} \delta g(\delta) d\delta \text{ and } \mathsf{pr}^L \geq \gamma^M \cdot (z^M - z^L)$$

to obtain

$$\mathsf{E}^L\delta \leq \frac{(1 - G(1 - z^M)\mathsf{E}[\delta|\delta \geq 1 - z^M]}{1 + G(1 - z^M)} < \frac{z^M}{2}.$$

Details are below, where we prove existence of the threshold $z^L$; specifically, see equation (65).

**Lemma 5 (Monotonicity of $\mathsf{E}^L\delta$)** *Function $\mathsf{E}^L\delta$ increases in $z^L$ and in $z^M$:*
*(i) $\partial\mathsf{E}^L\delta/\partial z^L > 0$ and (ii) $\partial\mathsf{E}^L\delta/\partial z^M > 0$.*

*Proof of (i):* Differentiating $\mathsf{E}^L\delta$ with respect to $z^L$, we obtain

$$\frac{\partial \mathsf{E}^L\delta}{\partial z^L} = \frac{1}{\mathsf{pr}^L}\left[\frac{\partial \mathsf{num}(\mathsf{E}^L\delta)}{\partial z^L} - \mathsf{E}^L\delta\frac{\partial \mathsf{pr}^L}{\partial z^L}\right] = \frac{1}{\mathsf{pr}^L}\left(\gamma^L\mathsf{E}^L\delta - \beta^L\right).$$

Since $\mathsf{pr}^L \geq 0$ (with equality only at $z^L = z^M$), it suffices to show that $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L) > 0$ for all $z^L \in [0, z^M)$. We will show that $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L)$ is strictly decreasing in $z^L$ on $[0, z^M0$. The desired inequality then follows because $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L) = 0$ at $z^L = z^M$. Differentiating $\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L)$, we obtain

$$\frac{\partial(\mathsf{pr}^L(\gamma^L\mathsf{E}^L\delta - \beta^L))}{\partial z^L} \quad = \quad -\frac{1}{4}(1 - z^L + \mathsf{E}^L\delta)g(1 - z^L) < 0 \tag{62}$$

*Proof of (ii):* Differentiating $\mathsf{E}^L\delta$ with respect to $z^M$, we obtain

$$\frac{\partial \mathsf{E}^L\delta}{\partial z^M} = \frac{1}{\mathsf{pr}^L}\left[\frac{\partial \mathsf{num}(\mathsf{E}^L\delta)}{\partial z^M} - \mathsf{E}^L\delta\frac{\partial \mathsf{pr}^L}{\partial z^M}\right] = \frac{1}{\mathsf{pr}^L}\left(\beta^M - \gamma^M\mathsf{E}^L\delta\right).$$

The derivative is positive if $\beta^M\mathsf{pr}^L - \gamma^M\mathsf{num}(\mathsf{E}^L\delta) > 0$. Expanding this,

$$\beta^M\mathsf{pr}^L - \gamma^M\mathsf{num}(\mathsf{E}^L\delta) \;=\; \frac{1}{4}\int_{1-z^M}^{1-z^L}\delta(1-z^L-\delta)g(\delta)d\delta\cdot\left(\gamma^M+\beta^M\right) > 0 \qquad (63)$$

**Lemma 6 (MLRP Results)** *For a family of densities $g(\delta|\theta)$ that obeys MLRP in $\theta$, i.e. for $\theta_1 > \theta_2$, $g(\delta|\theta_1)/g(\delta|\theta_2)$ increases in $\delta$, (i) probability of a limit buy $\mathsf{pr}^L$ decreases in $\theta$, and (ii) expectation $\mathsf{E}^L\delta$ decreases in $\theta$.*

Proof: to be typeset. (It is obtained by direct computation, using the definition of the monotone likelihood ratio property).

Lemma A.2.1 implies that solution $z^L$ is largest for the uniform distribution of innovations $\delta_t$. Hence, if this solution is below $z^M/3$ then $z^L$ is below $z^M/3$ for any distribution $\bar{g}$. Results for the uniform distribution can be obtained by direct (numerical) computation.

**Existence of $z^L_*(z^M)$.** First, we establish that for any given $z^M$ there exists $z^L_*(z^M) \in [0, z^M]$ that solves the indifference condition for the marginal limit order buyer (50). To see this, observe that

- At $z^L = 0$, we have $\mathsf{E}^L\delta > 0 = z^L$, since

$$\mathsf{num}(\mathsf{E}^L\delta) = \frac{1}{4}\int_{1-z^M}^{1}\delta(\delta-(1-z^M-1))g(\delta)d\delta > 0;$$

- At $z^L = z^M$, we have $\mathsf{E}^L\delta > 0 = z^L$.

To see this, note that both, $\mathsf{pr}^L$ and $\mathsf{num}(\mathsf{E}^L\delta)$ are 0 at $z^L = z^M$. Hence,

$$\begin{aligned}\mathsf{E}^L\delta|_{z^L=ym} &= \frac{\partial\mathsf{num}(\mathsf{E}^L\delta)/\partial z^L\mid_{z^L=z^M}}{\partial\mathsf{pr}^L/\partial z^L\mid_{z^L=z^M}} = \frac{(1/4)\cdot\int_{1-z^M}^{1}\delta g(\delta)d\delta}{(1/4)+(1/4)\cdot G(1-z^M)}\\ &= \frac{(1-G(1-z^M)\mathsf{E}[\delta\mid\delta\geq 1-z^M]}{1+G(1-z^M)} \qquad (64)\\ &\leq \frac{(1-G(1-z^M)(2-z^M)/2}{1+G(1-z^M)} \leq \frac{z^M}{2},\end{aligned}$$

where the inequalities follow because the uniform distribution FOSD $G$ (hence, $\mathsf{E}[\delta \mid \delta \geq 1 - z^M] \leq (2 - z^M)/2$ and $G(1 - z^M) \geq 1 - z^M$.

- Existence then follows by continuity of $\mathsf{E}^L\delta$.

**Lemma 7 (Bounds on $z_*^L(z^M)$)** *Any $z^L$ that solves $z^L = \mathsf{E}^L\delta$ for a given $z^M$ must be below $z^M/2$.*

*Proof:* The lemma follows since (*i*) $\mathsf{E}^L\delta$ is increasing in $z^L$ and (*ii*) at $z^L = z^M$ we have $\mathsf{E}^L\delta < z^M/2$.

**Uniqueness of $z_*^L(z^M)$.** To show uniqueness, we will show that for a fixed $z^M$ function $z(z^L, z^M) = \mathsf{E}^L\delta - z^L$ only crosses 0 once on $[0, z^M]$. Note that $z(0) > 0 > z(z^M)$. Since $z(\cdot)$ is continuous, it suffices to show that at $z^L$ such that $z(z^L) = 0$, we have $\partial z/\partial z^L < 0$. (That is $z(\cdot)$ must cross 0 from above and cannot touch the $x$-axis).

- We need to show that at $z^L$ such that $z(z^L) = 0$ (in what follows "at solution"), we have $\partial \mathsf{E}^L\delta/\partial z^L < 1$.

- At solution, $\partial \mathsf{E}^L\delta/\partial z^L < 1 \Leftrightarrow \mathsf{pr}^L > \gamma^L \mathsf{E}^L\delta - \beta^L \Leftrightarrow \mathsf{pr}^L > \gamma^L z^L - \beta^L$. Note that

$$\gamma^L z^L - \beta^L = \frac{1}{2}z^L - \frac{1}{4}(1 - G(1 - z^L))(\mathsf{E}[\delta \mid \delta > 1 - z^L] + z^L)$$

  We thus need to show that

$$\mathsf{pr}^L > \frac{1}{2}z^L - \frac{1}{4}(1 - G(1 - z^L))(\mathsf{E}[\delta \mid \delta > 1 - z^L] + z^L). \tag{65}$$

- At at $z^L$ such that $z(z^L) = 0$, we have $z^L \cdot \mathsf{pr}^L = \mathsf{num}(\mathsf{E}^L\delta)$. Rewrite this as follows:

$$\begin{aligned} \frac{1}{2}(z^M - z^L)z^L &= \frac{1}{4}\int_{1-z^M}^{1-z^L} (\delta - z^L) \cdot (\delta - (1 - z^M)) \cdot g(\delta) d\delta \\ &+ \frac{1}{4}(z^M - z^L)(1 - G(1 - z^L))\mathsf{E}[\delta \mid \delta > 1 - z^L] + z^L] \end{aligned}$$

  Use the above to rewrite inequality (65) as follows.

$$\mathsf{pr}^L > -\frac{1}{4}\frac{1}{z^M - z^L}\int_{1-z^M}^{1-z^L} (\delta - z^L) \cdot (\delta - (1 - z^M)) \cdot g(\delta) d\delta \tag{66}$$

Next, write the probability $\mathsf{pr}^L$ explicitly and rewrite (66) as

$$\frac{1}{4}(z^M-z^L)+\frac{\mu}{4}(z^M-z^L)G(1-z^L)+\frac{1}{4}\int\limits_{1-z^M}^{1-z^L}(\frac{\delta-z^L}{z^M-z^L}-1)\cdot(\delta-(1-z^M))\cdot g(\delta)d\delta>0,$$

which is equivalent to

$$\left(\frac{1}{4}+G(1-z^L)\right)(z^M-z^L)+\frac{1}{4}\int\limits_{1-z^M}^{1-z^L}\frac{\delta-z^M}{z^M-z^L}\cdot(\delta-(1-z^M))\cdot g(\delta)d\delta>0, \quad (67)$$

The first term is always positive. The second term is positive for $z^M \leq 1/2$ (since $\delta > 1-z^M \geq z^M$). $\Rightarrow$ remains to prove the above inequality for $z^M > 1/2$. Denote the left-hand side of the above inequality by $\Delta^L$. Observe that $\Delta^L = 0$ at $z^L = z^M$ (the first term is 0, and the second is 0 by l'Hôpital's rule). It thus suffices to show that $\Delta^L$ decreases in $z^L$ on $[0, z^M]$. Compute the appropriate derivative:

$$\begin{aligned}\frac{\partial\Delta^L}{\partial z^L} \quad = \quad & -\frac{1}{2}+\frac{1}{4}\big(-g(1-z^L)(1-z^L-z^M) \\ & +\frac{1}{(z^M-z^L)^2}\int\limits_{1-z^M}^{1-z^L}(\delta^2-\delta+z^M(1-z^M))g(\delta)d\delta \\ & +(1-G(1-z^L))-(z^M-z^L)g(1-z^L)\big)\,.\end{aligned}$$

Since $\delta^2-\delta$ is minimized at $\delta = 1/2$, the upper bound on the integral term depends on $z^L$. There are three possibilities (for $z^L < z^M$ and $z^M < 1/2$):

$(i)$ For $z^L < 1-z^M < 1/2 < z^M$, we have, for $\delta \in [1-z^M, 1-z^L]$, $\delta^2-\delta < (1-z^L)^2-(1-z^L)$, and further, $1-z^L-z^M > 0$. Thus

$$\begin{aligned}\int\limits_{1-z^M}^{1-z^L}(\delta^2-\delta+z^M(1-z^M))g(\delta)d\delta \quad & < \quad \frac{(1-z^L-z^M)(z^M-z^L)}{(z^M-z^L)^2}\int\limits_{1-z^L}^{1-z^M}g(]\delta)d\delta \\ & < \quad (1-z^L-z^M)g(1-z^L).\end{aligned}$$

Consequently, since $1-G(1-z^L) < 1$,

$$\frac{\partial\Delta^L}{\partial z^L} \quad < \quad -\frac{1}{2}+\frac{1}{4}(1-G(1-z^L))-\frac{1}{4}(z^M-z^L)g(1-z^L) < 0.$$

*(ii)* For $1 - z^M < z^L < {}^1\!/_2 < z^M$, we have, for $\delta \in [1 - z^M, 1 - z^L]$, $\delta^2 - \delta < (1 - z^M)^2 - (1 - z^M)$, and further, $1 - z^L - z^M < 0$ and $2z^L - 1 < 0$. The integral term is then negative. Consequently, since $1 - G(1 - z^L) < 1$,

$$\frac{\partial \Delta^L}{\partial z^L} \; < \; -\frac{1}{2} + \frac{1}{4}(1 - G(1 - z^L)) + \frac{1}{4}(2z^L - 1)g(1 - z^L) < 0.$$

*(iii)* For $1 - z^M < {}^1\!/_2 < z^L < z^M$, we have, for $\delta \in [1 - z^M, 1 - z^L]$, $\delta^2 - \delta < (1 - z^M)^2 - (1 - z^M)$, and further, $1 - z^L - z^M < 0$ and $2z^L - 1 > 0$. The integral term is then negative, and we have

$$\frac{\partial \Delta^L}{\partial z^L} \; < \; -\frac{1}{4} - \frac{1}{4}G(1 - z^L) + \frac{1}{4}(2z^L - 1)g(1 - z^L).$$

Using the upper bound on $g$ from expressions (44), it remains to show that

$$-1 - G(1 - z^L) + (2z^L - 1)\frac{G(1 - z^L)}{1 - z^L} < 0.$$

The above inequality is true for all $z^L < {}^3\!/_4$, since:

$$-(1 - z^L) + (-1 + z^L + 2z^L - 1)G(1 - z^L) < 4z^L - 3 < 0.$$

- We have thus shown that function $z(z^L, z^M) = \mathsf{E}^L\delta - z^L$ only crosses 0 once on $[0, z^M]$.

This completes the argument on existence and uniqueness of $z^L$ that solves the indifference equation for the limit order buyer, for all $z^M \in [0, {}^3\!/_4]$.

### A.2.2 Step 2: Monotonicity of $z^L_*(z^M)$

To show that $z^L$ increases in $z^M$, it suffices to show that the partial derivative of $z(z^L, z^M)$ in $z^M$ is positive (the positive partial derivative implies that $z$, viewed as a function of $z^L$, will then necessarily cross 0 further to the right, since it crosses from above). This is equivalent to showing that $\partial \mathsf{E}^L\delta / \partial z^M > 0$, which in turn follows from (Lemma 5).

### A.2.3 Step 3: Existence of $z^M$

We need to show existence and uniqueness of $z^M$ that solves equation (51):

$$z^M - \mathsf{E}^M\delta = \mathsf{pr}^M(z^M - z^L_*(z^M)).$$

**Notation and Preliminary Properties** Similarly to $\mathsf{E}^L\delta$, we need to derive the expression for $\mathsf{E}^M\delta$; omit subscripts $t$. We are assuming that $z^M \in [0, {}^3\!/\!_4]$.

$$\mathsf{E}^M\delta \;=\; \frac{\int_{-1}^{1} d\delta \int_{-1}^{1} dy(\delta \cdot h^M(\delta, y \mid \text{LB}))}{\mathsf{pr}^M},$$

where function $h^M(\delta, y \mid \text{MB})$ is defined as follows:

$$h^M(\delta, y \mid \text{MB}) = \begin{cases} \frac{1}{2} \cdot f(\delta), \text{ if } \delta \in [z^M - 1, 1] \text{ and } y \in [z^M - \delta, 1]; \\ 0, \text{ otherwise.} \end{cases}$$

and $\mathsf{pr}^M = \mathsf{Pr}[\text{MB}]$ is given by $\mathsf{pr}^M = \int_{-1}^{1} d\delta \int\limits_{-1}^{1} dy(h^M(\delta, y \mid \text{MB}))$. One more piece of short-hand notation: $\mathsf{num}(\mathsf{E}^M\delta) := \mu \int_{-1}^{1} d\delta \int_{-1}^{1} dy(\delta \cdot h^M(\delta, y \mid \text{MB}))$. Using this notation, we then have $\mathsf{E}^M\delta = \mathsf{num}(\mathsf{E}^M\delta)/\mathsf{pr}^M$.

Substituting $h^M$ in, putting in appropriate integral bounds and expressing $f$ as a function of $g$, we express the probability of a market buy as follows:

$$\mathsf{pr}^M \;=\; \frac{1}{4}(1 - z^M) + \frac{1}{4}(1 - z^M)\; G(1 - z^M) + \frac{1}{4} \int\limits_{1-z^M}^{1} \delta g(\delta) d\delta \equiv \gamma^M \cdot (1 - z^M) + \beta^M,$$

where $\gamma^M$ and $\beta^M$ are defined accordingly (and the same as in Steps 1-2). Next, derive the expression for $\mathsf{num}(\mathsf{E}^M\delta)$:

$$\mathsf{num}(\mathsf{E}^M\delta) \;=\; \frac{1}{4} \int\limits_{0}^{1-z^M} 2\delta^2 g(\delta) d\delta + \frac{1}{4} \int\limits_{1-z^M}^{1} \delta(1 - z^M + \delta) g(\delta) d\delta$$

Taking derivatives, we obtain:

$$\frac{d\mathsf{pr}^M}{dz^M} = -\gamma^M < 0, \text{ and } \frac{d\mathsf{num}(\mathsf{E}^M\delta)}{dz^M} = -\beta^M < 0. \tag{68}$$

**Lemma 8 (Monotonicity of $\mathsf{E}^M\delta$)** *Expectation $\mathsf{E}^M\delta$ increases in $z^M$.*

*Proof:* Differentiating $\mathsf{E}^M\delta$ with respect to $z^M$, we obtain

$$\frac{\partial \mathsf{E}^M\delta}{\partial z^M} = \frac{1}{\mathsf{pr}^M}\left[\frac{\partial \mathsf{num}(\mathsf{E}^M\delta)}{\partial z^M} - \mathsf{E}^M\delta \frac{\partial \mathsf{pr}^M}{\partial z^M}\right] = \frac{1}{\mathsf{pr}^M}\left(\gamma^M \mathsf{E}^M\delta - \beta^M\right).$$

Since $\mathsf{pr}^M > 0$, it suffices to show that $\mathsf{pr}^M(\gamma^M \mathsf{E}^M\delta - \beta^M) < 0$ for all $z^M \in [0, 1]$ and $\mathsf{pr}^M(\gamma^M \mathsf{E}^M\delta - \beta^M) \geq 0$ at $z^M = 1$. Differentiating $\mathsf{pr}^M(\gamma^M \mathsf{E}^M\delta - \beta^M)$, we obtain

$$\frac{d(\mathsf{pr}^M(\gamma^M \mathsf{E}^M\delta - \beta^M))}{dz^M} = -\frac{1}{4}\mathsf{pr}^M(1 - z^M + \mathsf{E}^M\delta)g(1 - z^M) < 0$$

To show that $\mathsf{pr}^M(\gamma^M \mathsf{E}^M\delta - \beta^M) \geq 0$ at $z^M = 1$, we need to show that at $z^M = 1$, $\gamma^M \mathsf{num}(\mathsf{E}^M\delta) - \mathsf{pr}^M\beta^M \geq 0$. When $z^M = 1$, we have $\gamma^M = {}^1\!/\!_2$ and thus

$$\gamma^M \mathsf{num}(\mathsf{E}^M\delta) - \mathsf{pr}^M\beta^M = \frac{1}{2}\frac{1}{4}\int_0^1 \delta^2 g(\delta)d\delta - \left(\frac{1}{4}\int_0^1 \delta g(\delta)d\delta\right)^2 \geq \frac{1}{16}\left(\int_0^1 \delta g(\delta)d\delta\right)^2 > 0$$

**Lemma 9 (MLRP Results)** *For a family of densities $g(\delta \mid \theta)$ that obeys MLRP in $\theta$, i.e. for $\theta_1 > \theta_2$, $g(\delta \mid \theta_1)/g(\delta \mid \theta_2)$ increases in $\delta$, (i) probability of market buy $\mathsf{pr}^M$ decreases in $\theta$, and (ii) expectation $\mathsf{E}^M\delta$ decreases in $\theta$.*

*Proof:* To be typeset (the result follows by direct computation, utilizing properties of MLRP).

**Existence of $z^M$.** We need to show existence of $z^M$ that solves equation (51), i.e., we need

$$(1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta + \mathsf{pr}^M z_*^L(z^M)) = 0$$

Existence follows by continuity. At $z^M = 0$, the LHS $= -\mathsf{E}^M\delta < 0$ (the inequality is strict, because $\mathsf{num}(\mathsf{E}^M\delta) > 0$ and $\mathsf{pr}^M > 0$). At $z^M = {}^3\!/\!_4$, we have

$$\begin{aligned}&(1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta + \mathsf{pr}^M z_*^L(z^M)) > (1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta \\ &> (1 - \mathsf{pr}^M)z^M - \mathsf{E}^M\delta \, |_{\text{for uniform distribution } g} > 0.\end{aligned}$$

### A.2.4 Step 4: Optimality of the Threshold Strategies

The intuition for the optimality of the threshold strategies stems from competitive pricing and stationarity of investor decisions. An investor's deviation from one equilibrium action to another equilibrium action will not affect equilibrium bid and ask prices or probabilities of the future order submissions. Consequently, it is possible to show that the difference between a payoff to a market order and a payoff to a limit order at the equilibrium price to an investor with an aggregate valuation above $z^M$ is strictly greater than 0. (The formal argument is to be typeset).

**Out-Of-The-Equilibrium-Beliefs.** A more complex scenario arises when an investor deviates from his equilibrium strategy by submitting an limit order at a price different to the prescribed competitive equilibrium price. Whether or not this investor expects to benefit from such a deviation depends on the reaction to this deviation by the low-latency liquidity providers and investors in the next period. For instance, can an investor increase the execution probability of his limit buy order by posting a price that is above the equilibrium bid price?

We employ a perfect Bayesian equilibrium concept. This concept prescribes that investors and low-latency liquidity providers update their beliefs by Bayes rule, whenever possible, but it does not place any restrictions on the beliefs of market participants when they encounter an out-of-equilibrium action.

To support competitive prices in equilibrium we assume that if a limit buy order is posted at a price different to the competitive equilibrium bid price $\mathsf{bid}^*_{t+1}$, then market participants hold the following beliefs regarding this investor's knowledge of the period $t$ innovation $\delta_t$.

- If a limit buy order is posted at a price $\widehat{\mathsf{bid}} < \mathsf{bid}^*_{t+1}$, then market participants assume that this investor followed the equilibrium threshold strategy, but "made a mistake" when pricing his orders. A low-latency liquidity provider then updates his expectation about $\delta_t$ to the equilibrium value and posts a buy limit order at $\mathsf{bid}^*_{t+1}$. The original investor's limit order then executes with zero probability.

- If a limit buy order is posted at a price above $\widehat{\mathsf{bid}} > \mathsf{bid}^*_{t+1}$, then market participants believe the this order stems from an investor from a sufficiently high aggregate valuation (e.g., $z_t = 2$) and update their expectations about $\delta_t$ to $\mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}]$ accordingly (to $\mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}] = 1$ if the belief on $z_t$ is $z_t = 2$). The new posterior expectation of $V_t$ equals to $p_{t-1} + \mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}]$. A low-latency liquidity provider is then willing to post a competitive bid price $\mathsf{bid}^{**}_{t+1} = p_{t-1} + \mathsf{E}[\delta_t \mid \widehat{\mathsf{bid}}] + \mathsf{E}[\delta_{t+1} \mid \mathrm{MS}_{t+1}]$. With the out-of-the-equilibrium belief of $z_t = 2$, a limit order with the new price $\mathsf{bid}^{**}_{t+1}$ outbids any limit buy order that yields investors positive expected profits.

The beliefs upon an out-of-equilibrium sell order are symmetric. The above out-of-equilibrium beliefs ensure that no investor deviates from his equilibrium strategy.

We want to emphasize that these beliefs and actions do *not* materialize in equilibrium. Instead, they can be loosely thought of as a "threat" to ensure that investors do not deviate from their prescribed equilibrium strategies.

# References

Anand, Amber, Tim McCormick, and Laura Serban, 2012, Does the make-take structure dominate the traditional structure? evidence from the options markets, Discussion paper, Syracuse University.

Angel, James, Lawrence Harris, and Chester Spatt, 2011, Equity trading in the 21st century, *The Quarterly Journal of Finance* 1, 1–53.

Battalio, Robert, and Craig Holden, 2001, A simple model of payment for order flow, internalization, and total trading cost, *Journal of Financial Markets* 4, 33–71.

Battalio, Robert H., Andriy Shkilko, and Robert A. Van Ness, 2012, To pay or be paid? The impact of taker fees and order flow inducements on trading costs in U.S. options markets, *SSRN eLibrary.*

Bessembinder, Hendrik, Jia Hao, and Michael L. Lemmon, 2012, Why designate market makers? Affirmative obligations and market quality, *SSRN eLibrary.*

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2012, Equilibrium high-frequency trading, *SSRN eLibrary.*

Colliard, Jean-Edouard, and Thierry Foucault, 2012, Trading fees and efficiency in limit order markets, *Review of Financial Studies, Forthcoming.*

Degryse, Hans, Mark Van Achter, and Gunther Wuyts, 2009, Dynamic order submission strategies with competition between a dealer market and a crossing network, *Journal of Financial Economics* 91, 319 – 338.

Foucault, T., 1999, Order flow composition and trading costs in a dynamic limit order market, *Journal of Financial Markets* 2, 99–134.

——— , O. Kadan, and E. Kandel, 2005, Limit order book as a market for liquidity, *Review of Financial Studies* 18, 1171–1217.

Foucault, Thierry, Ohad Kadan, and Eugene Kandel, 2012, Liquidity cycles and make/take fees in electronic market, *Journal of Finance, Forthcoming.*

Glosten, L., 1994, Is the electronic open limit order book inevitable?, *The Journal of Finance* 49, 1127–1161.

——— , and P. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogenously informed traders, *Journal of Financial Economics* 14, 71–100.

Goettler, R., C. Parlour, and U. Rajan, 2005, Equilibrium in a dynamic limit order market, *Journal of Finance* 60, 2149–2192.

———, 2009, Informed traders and limit order markets, *Journal of Financial Economics* 93, 67–87.

Hasbrouck, Joel, and Gideon Saar, 2011, Low-latency trading, *SSRN eLibrary.*

Hendershott, T., C. Jones, and A. Menkveld, 2011, Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1–33.

Hendershott, Terrence J., and Ryan Riordan, 2012, Does algorithmic trading improve liquidity?, *Journal of Financial and Quantitative Analysis, Forthcoming.*

Hoffmann, Peter, 2012, A dynamic limit order market with fast and slow traders, *SSRN eLibrary.*

Jovanovic, Boyan, and Albert J. Menkveld, 2011, Middlemen in limit-order markets, *SSRN eLibrary.*

Kandel, Eugene, and Leslie M. Marx, 1999, Payments for order flow on Nasdaq, *Journal of Finance* 49, 35–66.

Kaniel, R., and H. Liu, 2006, So what orders so informed traders use?, *Journal of Business* 79, 1867–1913.

Malinova, Katya, and Andreas Park, 2011, Subsidizing liquidity: The impact of make/take fees on market quality, *SSRN eLibrary.*

McInish, Thomas H., and James Upson, 2012, Strategic liquidity supply in a market with fast and slow traders, *SSRN eLibrary.*

Milgrom, P., and N. Stokey, 1982, Information, trade and common knowledge, *Journal of Economic Theory* 26, 17–27.

Parlour, C., 1998, Price dynamics in limit order markets, *Review of Financial Studies* 11, 789–816.

Parlour, Christine, and Uday Rajan, 2003, Payment for order flow, *Journal of Financial Economics* 68, 379–411.

Parlour, C., and D. Seppi, 2008, Limit order markets: A survey, in A. W. A. Boot, and A. V. Thakor, ed.: *Handbook of Financial Intermediation and Banking* (Elsevier Science).

Rosu, I., 2009, A dynamic model of the limit order book, *Review of Financial Studies* 22, 4601–4641.

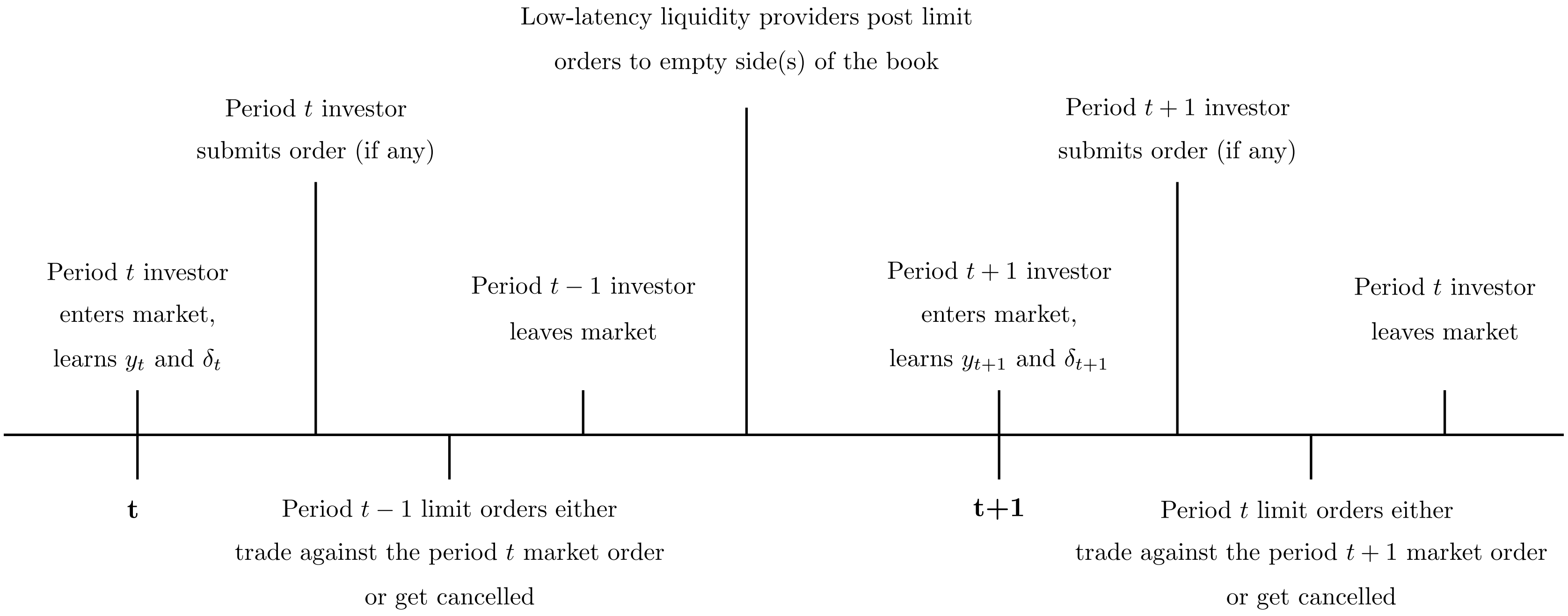
——— , 2011, Liquidity and information in order driven markets, Discussion paper, HEC.

Securities and Exchange Commission, 2010, Concept release on market structure, Release No. 34-61358, File No. S7-02-10, Discussion paper, Securities and Exchange Commission http://www.sec.gov/rules/concept/2010/34-61358.pdf.

Skjeltorp, Johannes A., Elvira Sojli, and Wing Wah Tham, 2012, Identifying cross-sided liquidity externalities, *SSRN eLibrary.*

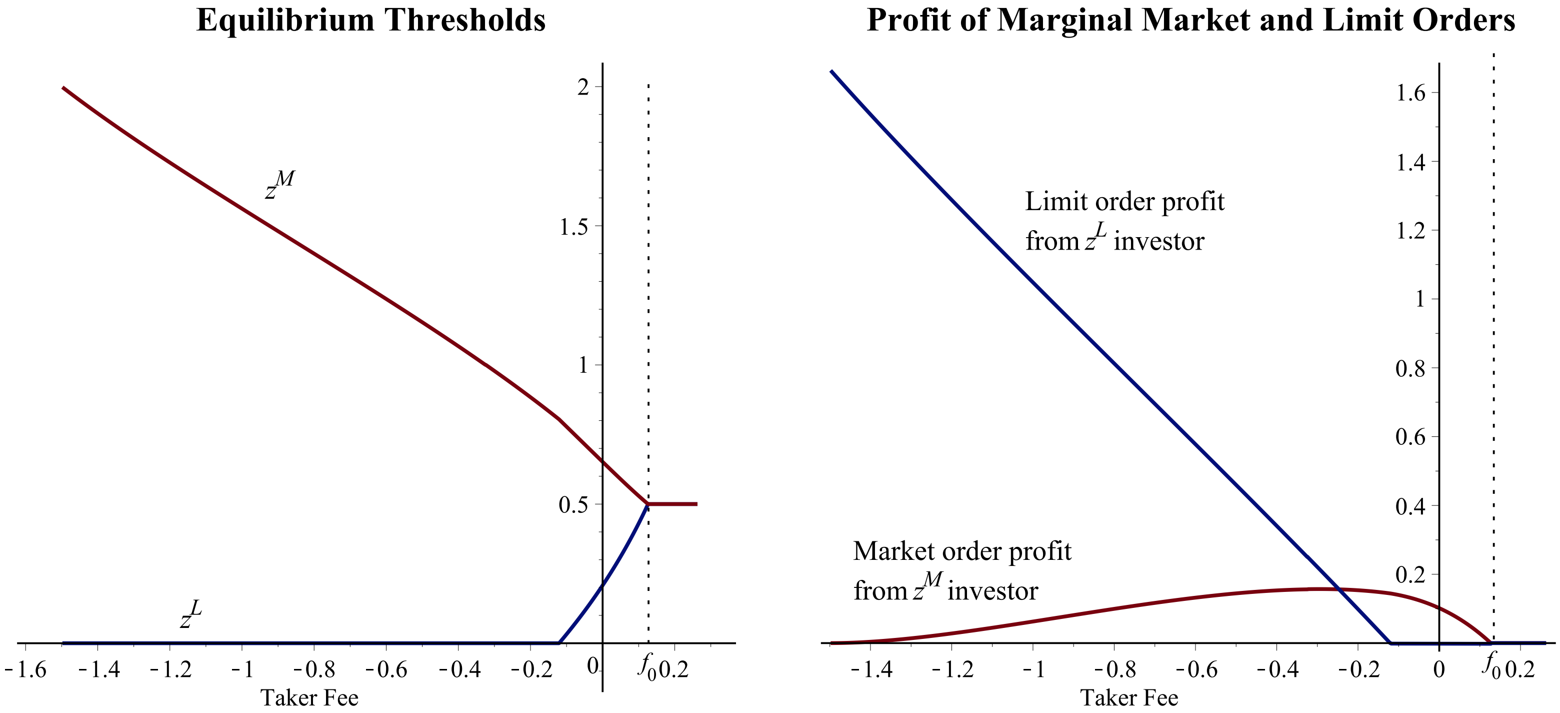
## Figure 1: Entry and Order Submission Timeline

This figure illustrates the timing of events upon the arrival of an investor at an arbitrary period, $t$, until their departure from the market. Value $y_t$ is the private valuation of the period $t$ investor and $\delta_t$ is the innovation to the security's fundamental value in period $t$.


Low-latency liquidity providers post limit orders to empty side(s) of the book
Period $t$ investor submits order (if any)
Period $t+1$ investor submits order (if any)
Period $t$ investor enters market, learns $y_t$ and $\delta_t$
Period $t-1$ investor leaves market
Period $t+1$ investor enters market, learns $y_{t+1}$ and $\delta_{t+1}$
Period $t$ investor leaves market
t
Period $t-1$ limit orders either trade against the period $t$ market order or get cancelled
t+1
Period $t$ limit orders either trade against the period $t+1$ market order or get cancelled

### Figure 2: Equilibrium Thresholds and Payoffs to the Marginal Market and Limit Orders: Flat Fee Model

The left panel depicts the equilibrium aggregate valuations $z^M$ (red line) and $z^L$ (blue line) for the marginal market and limit order submitters, respectively. The right panel depicts the expected payoff that the investors with an aggregate valuation of $z^M$ and $z^L$ receive in equilibrium, as functions of the taker fee $f$. Both panels are for the setting where investors pay a flat, average fee per trade. An investor submits a market buy order when his aggregate valuation $z_t$ is above $z^M$, a limit buy order when $z^L \leq z_t < z^M$, and abstains from trading when $|z_t| < z^L$; sell decision are symmetric to buy decisions. The plot illustrates that as $f$ increases, investors submit more market orders and fewer limit orders. There exist the level of the taker fee, $f^{NT} = -1.5 < 0$ and $f_0 > 0$, at which the investor with aggregate valuation $z^M$ receives zero profit from submitting a market order. The plot illustrates that investors only submit limit orders for values of $f < f_0$. Parameter $\alpha$ in the distribution of innovations is set to $\alpha = 1$; results for other values of $\alpha$ are qualitatively similar.



# Figure 3: Trading Volume and Market Participation: Flat Fee Model

The left panel plots trading volume, measured as $\mathsf{Pr}$(market order), as a function of the taker fee $f$, for the setting where investors pay a flat fee per trade. The right panel plots the level of market participation, measured as $\mathsf{Pr}$(market order) + $\mathsf{Pr}$(limit order), as a function of the taker fee level $f$. The value $f_0$ represents the taker fee level at which the equilibrium threshold values $z^M$ and $z^L$ coincide, and the marginal market order submitter $z^M$ earns zero profits in expectation. Parameter $\alpha$ in the distribution of innovations is set to $\alpha = 1$; results for other values of $\alpha$ are qualitatively similar.

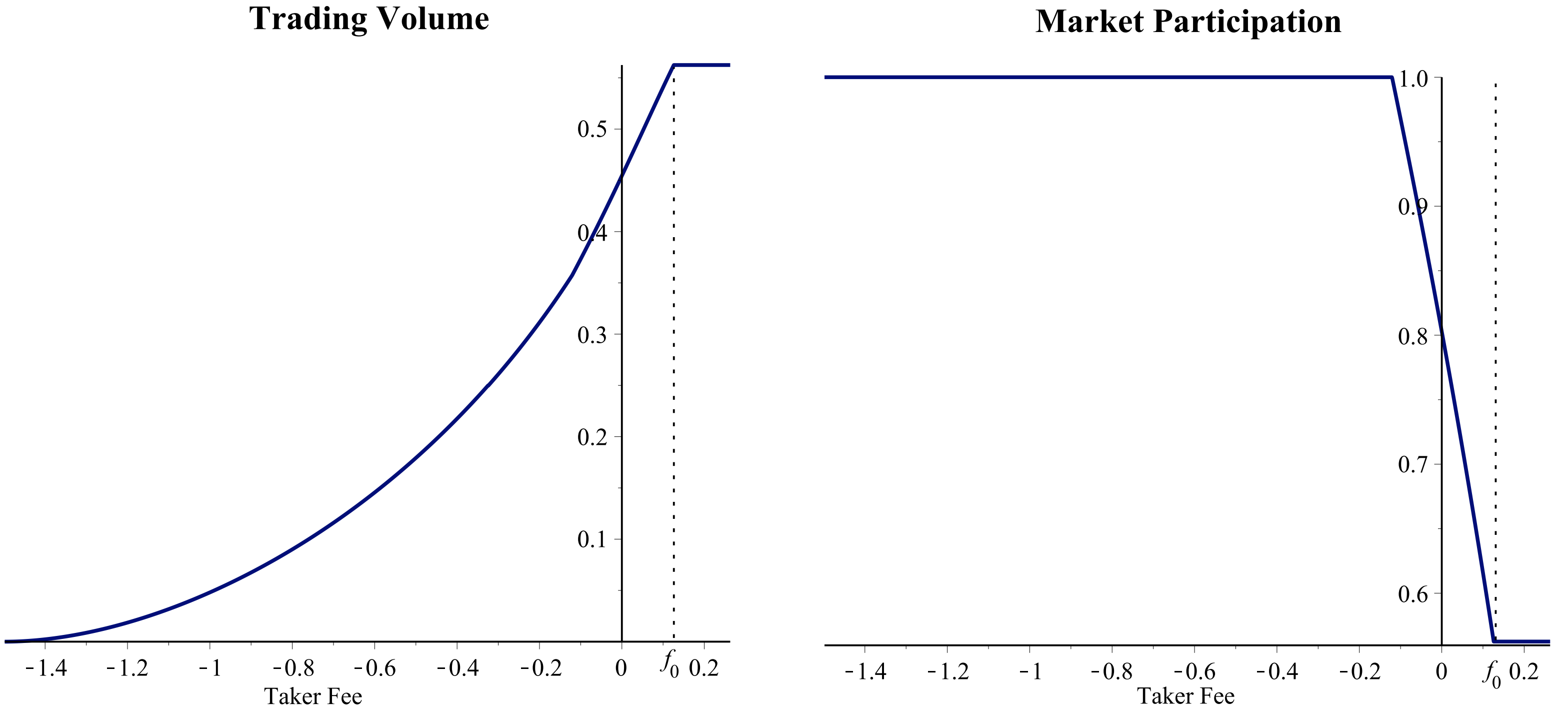
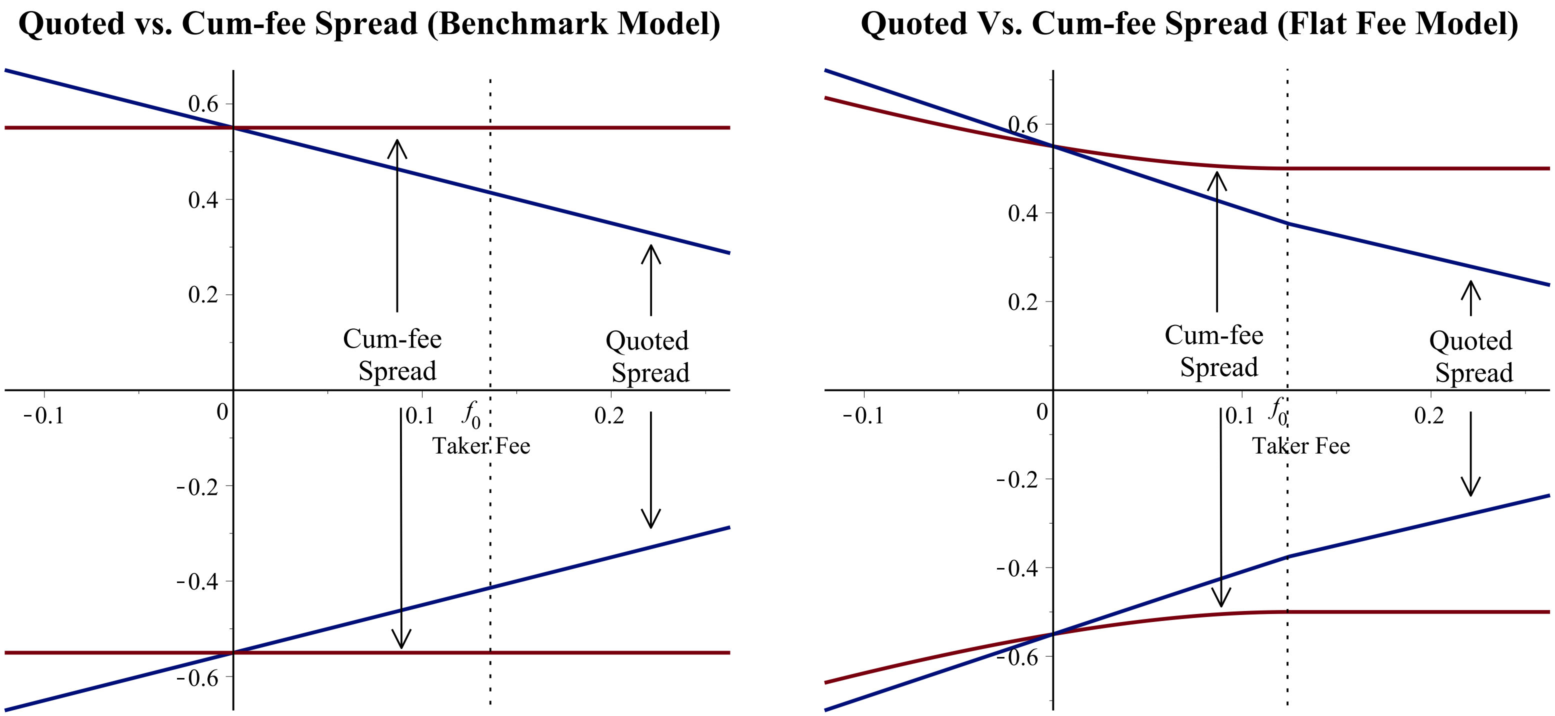


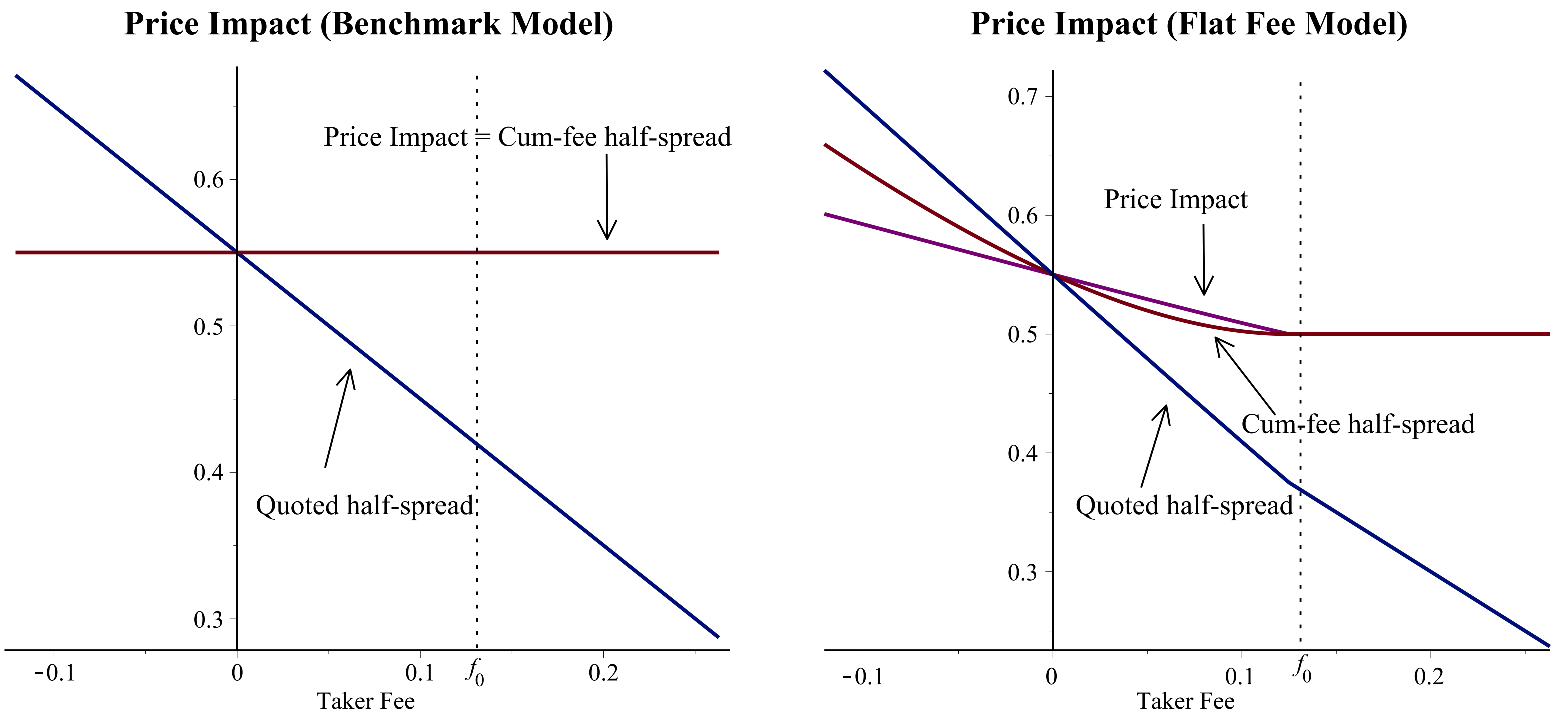
**Figure 4: Quoted and Cum-Fee Spreads**

The left panel plots the quoted spread (the inner, blue lines) and the cum-fee spread (the outer, red lines) as a function of the taker fee $f$, for the benchmark setting. The right panel plots the quoted spread (the inner, blue lines) and the cum-fee spread (the outer, red lines) as a function of the taker fee $f$, for the setting in which the investor pays a flat fee per trade. The value $f_0$ represents the taker fee level at which the equilibrium threshold values $z^M$ and $z^L$ coincide, and the marginal market order submitter $z^M$ earns zero profits in expectation. Parameter $\alpha$ in the distribution of innovations is set to $\alpha = 1$; results for other values of $\alpha$ are qualitatively similar.



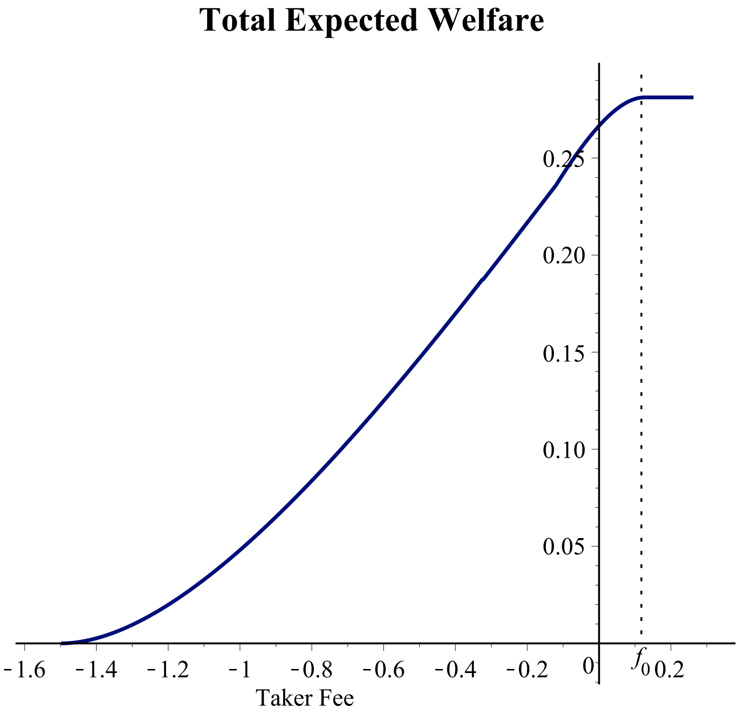
### Figure 5: Price Impact

The left panel plots price impact, quoted, and cum-fee half-spreads as functions of the taker fee $f$ for the benchmark setting where investors pay exchange maker-taker fees per trade. The right panel plots price impact, quoted, and cum-fee half-spreads as functions of the taker fee $f$ for the setting in which investors pay a flat fee per trade. The value $f_0$ represents the taker fee level at which the equilibrium threshold values $z^M$ and $z^L$ coincide, and the marginal market order submitter $z^M$ earns zero profits in expectation. Parameter $\alpha$ in the distribution of innovations is set to $\alpha = 1$; results for other values of $\alpha$ are qualitatively similar.



## Figure 6: Social Welfare: Flat Fee Model

The figure plots total expected social welfare, as defined in Section 5, as a function of the taker fee $f$, for the setting where investors pay a flat fee per trade. The value $f_0$ represents the taker fee level at which the equilibrium threshold values $z^M$ and $z^L$ coincide, and the marginal market order submitter $z^M$ earns zero profits in expectation. Parameter $\alpha$ in the distribution of innovations is set to $\alpha = 1$; results for other values of $\alpha$ are qualitatively similar.


Total Expected Welfare
$f_0$
Taker Fee