Roel Oomen

Last look

Article (Accepted version) (Refereed)

Original citation:
Oomen, Roel (2016) *Last look.* Quantitative Finance. pp. 1-14. ISSN 1469-7688

DOI: 10.1080/14697688.2016.1262545

Last Look

Roel Oomen*

November, 2016

Abstract

In over-the-counter markets, a trader typically sources indicative quotes from a number of competing liquidity providers, and then sends a deal request on the best available price for consideration by the originating liquidity provider. Due to the communication and processing latencies involved in this negotiation, and in a continuously evolving market, the price may have moved by the time the liquidity provider considers the trader's request. At what point has the price moved too far away from the quote originally shown for the liquidity provider to reject the deal request? Or perhaps the request can still be accepted but only on a revised rate? "Last look" is the process that makes this decision, i.e. it determines whether to accept – and if so at what rate – or reject a trader's deal request subject to the constraints of an agreed trading protocol. In this paper I study how the execution risk and transaction costs faced by the trader are influenced by the last look logic and choice of trading protocol. I distinguish between various "symmetric" and "asymmetric" last look designs and consider trading protocols that differ on whether, and if so to what extent, price improvements and slippage can be passed on to the trader. All this is done within a unified framework that allows for a detailed comparative analysis. I present two main findings. Firstly, the choice of last look design and trading protocol determines the degree of execution risk inherent in the process, but the effective transaction costs borne by the trader need not be affected by it. Secondly, when a trader adversely selects the liquidity provider she chooses to deal with, the distinction between the different symmetric and asymmetric last look designs fades and the primary driver of execution risk is the choice of trading protocol.

*Roel Oomen is employed as the global co-head of electronic FX spot trading at Deutsche Bank A.G. The views and opinions rendered in this paper reflect the author's personal views about the subject and do not necessarily represent the views of Deutsche Bank AG or any part thereof. This article is necessarily general and is not intended to be comprehensive, nor does it constitute legal or financial advice in relation to any particular situation. Oomen would like to thank two anonymous referees, Andrew Chaplin, Martin Evans, Thierry Foucault, Ankur Pruthi, Eric Zijdenbos, colleagues at Deutsche Bank, and the participants at the "Microstructure of Foreign Exchange Markets" conference at the Cambridge-INET Institute, TradeTech FX 2016 in London, and the Fields Institute in Toronto for helpful comments.

1 Introduction

When conducting trade, the typical protocol is one where on the basis of advertised prices and quantities, the buyer extends an offer to the seller, who after consideration will then either accept or reject the offer thereby concluding the contract formation process. It is the seller – and not the buyer – that has the final say and determines whether or not a contract is formed (see, e.g., Part II of UNCITRAL, 2010). This protocol is closely mirrored in many of the over-the-counter financial markets where liquidity providers (i.e. the sellers of liquidity) publish indicative bid and offer prices at which they are willing to deal in a specified asset on a bi-lateral and disclosed basis with a trader that demands this liquidity (i.e. the buyer of liquidity). The trader can make an offer to deal on the basis of the indicated prices which the liquidity provider will then either accept or reject, i.e. the liquidity provider owns finality of the deal. The mechanism adopted by the liquidity provider to determine whether to accept or reject the deal request is commonly referred to as the "trade acceptance" or "last look" process.[1] In recent years, there has been a growing interest in last look with a particular focus on the role it plays in the spot foreign exchange markets (see, e.g., Bank of England, H.M. Treasury, and Financial Conduct Authority, 2015; Norges Bank Investment Management, 2015; Cartea, Jaimungal, and Walton, 2015). Does last look (continue to) serve a legitimate purpose? Does it give liquidity providers an unfair advantage over the liquidity takers? Are certain last look designs superior to others? The objective of this paper is to contribute to this debate by formulating a framework that can be used to address these types of questions and to provide insights into the key trade-offs and considerations faced when it comes to last look design.

The process of trading takes time: it is not instantaneous, not even in the world's largest financial market of foreign exchange where most trading is conducted via electronic channels without human intervention. The counterparts to any bi- or multi-lateral trade are necessarily in different physical locations and the communication required to facilitate the trading process involves transmission latencies. Additionally, decision making by the trader and the liquidity provider involves processing latencies, and in the case where a human is involved, reaction latencies are added.[2] At the same time, financial markets and information flows naturally evolve in continuous time and it is

[1]The term "last look" is also used in financial markets to refer to situations where a buyer gives the seller a final opportunity to revise its quote to match or improve on that which it has secured from a different counterpart. That process is entirely separate and not related in any way to the electronic trading last look workflows detailed in this paper.

[2]Human reaction speed is measured at around 150 – 200 milliseconds by Amano, Goda, Nishida, Ejima, Takeda, and Ohtani (2006). One-way transmission latencies tend to be in microseconds within the same data centre, sub-millisecond within the same city, around 10 milliseconds between Chicago and New York (e.g. http://www.mckay-brothers.com), 30 milliseconds between New York and London (e.g. http://www.hibernianetworks.com), and over 100 milliseconds between London and the Asia Pacific. Processing latencies can be as low as single digit micro-seconds, even for fairly complex tasks, see, e.g. Cliff, Brown, and Treleaven (2010).

therefore inevitable that from the point where the liquidity provider sent the indicative price to the point where it receives back a trader's deal request, the price, or the available liquidity may have changed. This is what necessitates a last look mechanism: it is needed to determine whether incoming offers to deal can be filled. It is the over-the-counter equivalent of an exchange's matching engine. In practice, the scope of the last look mechanism is much broader than the monitoring of intervening price movements and typically involves a number of fully automated validation checks (such as credit utilisation, trader permissions, message integrity, system health). Additionally, it serves as a risk management tool that lets the liquidity provider control its inventory in real-time (e.g. via automated circuit breakers) whilst streaming quotes to many individual traders simultaneously thereby offering out aggregate liquidity that far exceeds its risk absorbing capacity if all traders were to demand it at the same time. For the purposes of this paper, I restrict attention to the last look component that establishes whether a deal request can be accepted on the basis of the price information at the point of decision.

The last look process may include a hold-period or so-called "latency buffer" which postpones the trade acceptance decision for a specified period of time (typically measured in milli-seconds). This acts as a defensive measure on behalf of the liquidity provider to mitigate adverse selection costs that arise from a trader's latency sensitive flow. In a fragmented market where traders and liquidity providers are dispersed, with major trading platforms and news agencies located in different financial centres, and with market data distributed via venue specific publication protocols, it is a substantial technological and intellectual challenge for the liquidity provider to gather all the required information in a timely manner and to produce and deliver prices to the trader that accurately reflect the current state of the market on a continuous basis. At the same time, a trader tends to put multiple liquidity providers in competition for her flow. She may use an aggregator to consolidate the liquidity that is provided and send deal requests on the best price available. Clearly, if a liquidity provider is not the first to update its prices following a market move, then either its bid or offer price will stand out in the aggregator in comparison to some or all of its competitors' prices and appear attractive – albeit artificially so – for the trader to deal on. This is a mechanism by which the liquidity provider gets adversely selected: it does not receive a random or representative subset of the trader's deal requests, but it is more likely to receive interest when prices are stale or otherwise mis-priced.[3] This is a manifestation of an information asymmetry that is skewed in the trader's favour. Each liquidity provider is blind to the prices its competitors are showing, but the trader can observe and aggregate all the information and adverse selection is introduced merely by the act of price comparison and the latencies involved in the process of trading and price

[3] A nice illustration of such dynamics is found in the literature on the so-called "SOES bandits". These traders operated scalping strategies that systematically attempted to exploit stale firm quotes posted by the Nasdaq's dealers on the Small Order Execution System. See Harris and Schultz (1998), Foucault, Röell, and Sandås (2003) and references therein.

discovery. So even when the trader's flow is entirely uninformed (meaning that it is not motivated by anticipated short-term price movements over horizons relevant to the execution), once it has passed through the aggregator and reaches the liquidity provider, each will perceive it as adverse latency sensitive flow. Of course, this effect is further amplified when the trader's flow is in fact driven by short-term price predictions, or aggressive sweep-style execution techniques are adopted (see Oomen, 2016). An effective way for the liquidity provider to defend itself against this is via the inclusion of a latency buffer in the last look process. During this brief pause, the liquidity provider can ensure all its information sources are up-to-date and prevent the completion of trades at excessively stale or incorrect rates. As noted by Budish, Cramton, and Shim (2015), "*The high-frequency trading arms race is a symptom of flawed market design.*" The latency buffer[4] is designed to avoid precisely such a situation. It facilitates a well functioning market where counter-parties' trading activity is motivated by genuine need and a willingness to share risk rather than by opportunistic latency motivated reasons.

In this paper I use the model introduced by Oomen (2016), where multiple liquidity providers compete for a trader's uninformed flow, and I study how the process of trading is impacted by the choice of last look design. In its simplest form, last look specifies a tolerance level to an adverse price movement over the latency buffer which, when exceeded, makes the liquidity provider reject the deal request. I refer to this as one-sided last look – others use the term "asymmetric last look" to highlight the feature that deal requests never get rejected due to a price move that is in the liquidity provider's favour. Two-sided last look is a natural extension where deal requests are rejected on the basis of excessive price moves in either direction. This has loosely been referred to as "symmetric last look", although a precise definition of what constitutes symmetry has been lacking. The first contribution this paper makes is to unambiguously define the different last look designs and make explicit how each alters the execution risk faced by the trader. For example, I analyse four distinct two-sided last look designs and show that while each leads to materially different reject rates, all can be viewed as symmetric by specific, well-defined criteria. I also study a one-sided last look methodology and a two-sided model that provides rebates to the trader. All this is done within a unified framework that allows for a detailed comparative analysis.

Last look is applied in the broader context of a specified trading protocol, i.e. a set of rules that governs the possible outcomes of the last look process. I distinguish between two fundamentally different protocols; (i) where the deal request, if not rejected, can only be accepted at the indicative rate contained in the request and (ii) where the deal request, if not rejected, can be accepted at a different rate – for instance at a rate that incorporates any price

[4]Conceptually similar mechanisms have recently been deployed on a number of the major inter-dealer foreign exchange platforms (e.g. ParFX, EBS, Reuters) as well as on selected US equity platform (e.g. IEX, Chicago Stock Exchange).

improvement or slippage associated with intervening price moves over the latency buffer.[5] I show that a trader can eliminate any uncertainty of rate, uncertainty of execution, or both through the choice of trading protocol and last look configuration.

The main contribution of this paper is a comprehensive and analytical characterisation of the effective transaction costs and constituent components that are borne by the trader. An important finding is that the same transaction costs can be attained for any combination of last look design and trading protocol. This contradicts the notion commonly quoted that asymmetric last look is "unfair" and symmetric last look is somehow "fair". I argue that the choice of last look design is not an issue of fairness, but one of preference over liquidity access, trading style, and tolerance to execution risk. A further finding is that when the trader's flow imposes a high degree of adverse selection on the liquidity providers (e.g. as a result of aggregation, latency sensitive trading, or aggressive sweep-style execution), the distinction between the different symmetric and asymmetric last look designs fades and is of second order importance compared to the choice of trading protocol.

The remainder of the paper is organised as follows. Section 2 lays out the model framework which is then used in Section 3 to define and study a number of differing last look designs and trading protocols. Section 4 extends this analysis by focussing on the effective transaction costs incurred by the trader for each scenario. Section 5 concludes and the appendix contains the proofs.

2 The model

I adopt the same setup as in Oomen (2016). There are N liquidity providers (LPs) that compete for a trader's uninformed flow. The trader is assumed to require liquidity for exogenously motivated reasons (a noise trader as in Kyle, 1985) and will direct a deal request to buy or sell a specified asset to the LP that shows the best price at that point in time, i.e. she will approach the LP that is top-of-book in the aggregator on the side she wants to deal. The LPs are characterised by their individual ability to discover the unobserved true price process, and charge a spread around their best estimate of the true mid price. Upon winning a deal request from the trader, the LP needs to decide whether to accept – and if so at what rate – or reject the deal request. The LP does so by applying a last look rule. The liquidity originates from the LP and it therefore owns finality of the deal, just as the matching engine of an exchange determines whether an order is filled or not. But different from anonymous trading on-exchange, the model here reflects an over-the-counter (OTC) market structure where (i) counter-parties are known to each other and trading takes place on a disclosed and bi-lateral basis, (ii) each LP is blind to the liquidity that the competing

[5] The trading protocol may also allow for acceptance in partial amount. This is a relevant scenario, but not one that I study in this paper.

LPs are providing to the same trader, i.e. they do not observe a centralised consolidated order book prior to sending a quote, and (iii) the liquidity provided is bespoke to the trader, taking the nature of her flow and aggregator setup into account.

The unobserved true (logarithmic) price process is assumed to follow a random walk:

$$p_t^* = p_{t-1}^* + \varepsilon_t, \tag{1}$$

with $\varepsilon \sim$ i.i.d. $\mathcal{N}(0,\sigma^2)$. Each LP produces independently of one another an estimate, $p^{(i)}$, of the true price process and incurs a measurement error, $m^{(i)}$, in doing so:

$$p_t^{(i)} = p_t^* + m_t^{(i)}, \tag{2}$$

$$m_t^{(i)} = \beta_i m_{t-1}^{(i)} + \eta_t^{(i)}, \tag{3}$$

for $i \in \{1,2,\dots,N\}$, $\eta^{(i)} \sim$ i.i.d. $\mathcal{N}(0,(1-\beta_i^2)\omega_i^2)$, $0 \le \beta < 1$, and $\text{corr}(\eta_t^{(i)},\eta_t^{(j)}) = \rho_{i,j}$ for $i \neq j$. An increase in ω decreases the accuracy of the estimate whereas an increase in β increases the persistence of the measurement error. An LP with superior price discovery skills is therefore characterised by small ω and β parameters. While the LPs act independently, their information sets may overlap or their methodologies may share common features. This can lead to cross-sectional correlation in their measurement error and the parameter ρ captures this effect. Each LP posts a bid price (b) and an offer price (a) at which it is willing to buy and sell the asset in a standard amount by charging a nominal spread $s = a - b$ centred around its best estimate of where the true price is at any point in time:

$$b_t^{(i)} = p_t^{(i)} - \frac{s_i}{2} \qquad \text{and} \qquad a_t^{(i)} = p_t^{(i)} + \frac{s_i}{2}. \tag{4}$$

While the model allows for heterogeneity of the LPs, where each is characterised by a unique set of parameters (see Oomen, 2016, for an analysis of such a setup), in this paper I assume for simplicity that the LPs are homogenous with identical parameters (ω,β,ρ,s). Note, however, this does not mean the LPs bid and offer prices will be the same. Homogeneity merely implies that the processes driving the price formation of each LP are governed by a uniform set of parameters.

The trader uses an aggregator to consolidate the liquidity provided by the LPs and is then assumed to trade following the rule below.

Definition 1 (best price routing) *If the trader wants to buy, she will attempt to do so at price* $\underline{a}_t = \min_i a_t^{(i)}$ *by routing the deal request to the originating liquidity provider, LP-*$\underline{i}_t = \arg\min_i a_t^{(i)}$. *Analogously, if the trader wants to sell, she will attempt to do so at price* $\overline{b}_t = \max_i b_t^{(i)}$ *by routing the deal request to LP-*$\overline{i}_t = \arg\max_i b_t^{(i)}$.

I assume throughout that the liquidity provided by the LPs is sufficient to satisfy the trader's demand at any point in time and she therefore routes the full order to a single LP (stack-sweep execution, where the trader attempts to deal with multiple LPs simultaneously, is discussed in Oomen, 2016).

With disagreement amongst the LPs on where the true price process is, the observed inside spread in the aggregator (i.e. $\underline{a} - \overline{b}$) will always be tighter than the nominal spread charged by any individual LP. As shown in Oomen (2016), with N homogenous LPs, the trader can expect to observe a spread in the aggregator equal to:

$$S \equiv E(\underline{a}_t - \overline{b}_t) = s - 2\omega\sqrt{1-\rho}\,\psi_N, \tag{5}$$

where $\psi_N = E(\max_i\{u_i\}_{i=1}^N)$ for $u_i \sim$ i.i.d. $\mathcal{N}(0,1)$. Because ψ_N grows in N, there is no lower-bound to the observed spread and for sufficiently large N (or ω) it can be arbitrarily negative. Depending on the number of LPs participating in the aggregator and their characteristics, the trader may observe sustained inversion of the bid and offer prices. In the absence of last look, this would constitute a guaranteed arbitrage opportunity and therefore be a flawed and unsustainable setup. But as already discussed, the prices shown by the LP are indicative and any deal requests submitted by the trader on the basis of these prices may be accepted or rejected at the discretion of the LP. The spread observed in the aggregator is therefore not necessarily the spread at which the trader is able to transact due to the execution risk associated with the LPs' trade acceptance process. Meaningful transaction cost analysis should therefore not merely focus on the nominal spread charged by the individual LPs, or the observed spread in the aggregator, but instead measure the *effective* spread that takes full account of the execution risk inherent in the trade process. Oomen (2016) shows that the key determinants of the effective spread are (i) the aggregator setup, e.g. how many and which LPs are included, (ii) the trader's execution style, e.g. full-amount or stack-sweep, (iii) the LPs' liquidity offering, e.g. their prices, nominal spreads, and trade acceptance criteria, and (iv) the LPs' risk management approach, e.g. internalisation versus externalisation. Below I will expand on this study by analysing execution risk and effective transaction costs for different trading protocols and last look trade acceptance methodologies.

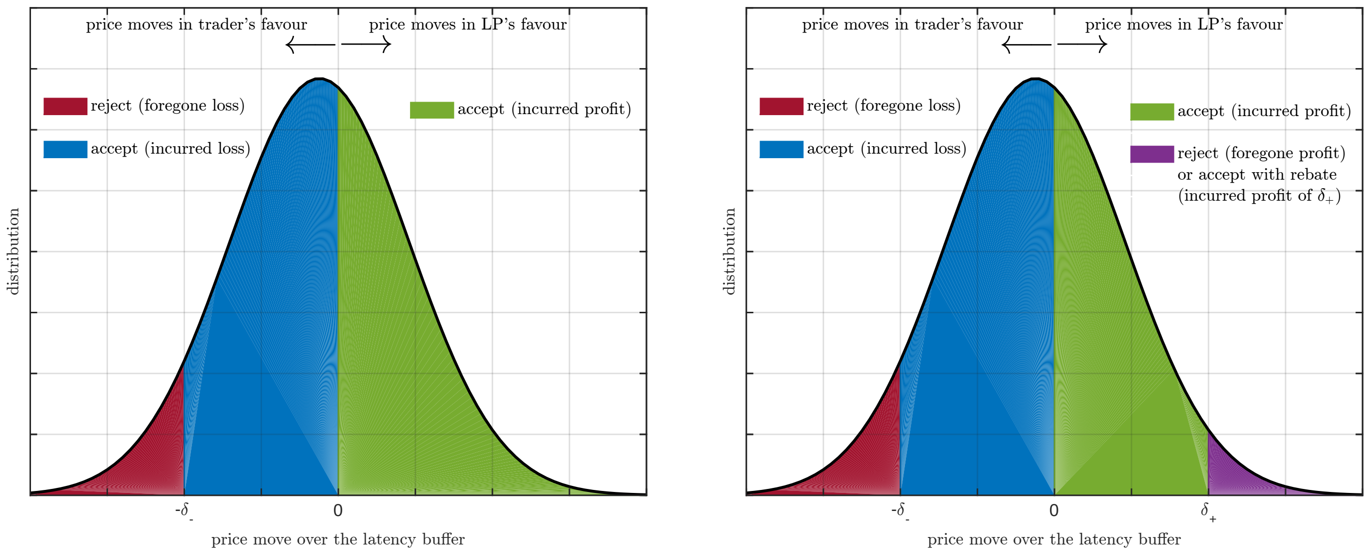
3 The trading protocol & last look design

Below I distinguish between two different trading protocols. The first is one where the LP supplies indicative prices, the trader makes a request to deal on those prices, and the LP then uses last look to decide whether to accept at that price or reject the request. It is a binary decision and does not allow for acceptance on a revised rate or in a partial amount. I refer to this protocol as trading on the indicative rate or, in short, as trading "on rate". It provides the trader with certainty of rate but not certainty of execution: it is similar to a fill-or-kill or a marketable limit order

Figure 1: Last look for on-rate trading

Panel A : on-rate with one-sided last look

Panel B : on-rate with two-sided last look



Note. This figure draws the distribution of the trade sign adjusted price move α (as defined in Eq. 7) over the latency buffer n immediately following a deal request from a trader that adopts best price routing as in Definition 1. The shaded areas indicate the various accept and reject (or rebate) regions associated with one- and two-sided last look for the on-rate trading protocol as in Definition 2 (or 3). The foregone and incurred profits and losses indicated in the chart are as seen from the LP's perspective (and opposite to those from the trader's perspective).

in on-exchange trading. The second trading protocol is one where the LP supplies an indicative rate, but the trader now makes a deal request referencing – but not insisting on – that price, and the LP subsequently accepts the deal request at the time of decision and the then current price. I refer to this protocol as trading "on-latest" reflecting that the rate is only determined immediately prior to completion of the trade. Relative to the indicative price that the trader observed when she sent the deal request, the final price incorporates any slippage or price improvement associated with the price variation over the intervening time interval. This protocol offers the trader certainty of execution but not certainty of rate. It is the OTC equivalent of an on-exchange market order.

3.1 Trading on the indicative rate

For the on-rate protocol, the last look rule adopted by the LP should establish whether a deal request made on an indicative price shown at time t is accepted or rejected by the originating LP at the point of decision at time $t + n$ on the basis of the intervening price movement. Specifically, at what point have prices deviated too far from the indicative price originally shown for the deal request to still be accepted? The time-interval n represents the accumulation of transmission, processing, and reaction latencies as well as the possible inclusion of an additional hold period or latency buffer. For ease of exposition, I equate n to the latency buffer whilst emphasising that the other latency components cannot be avoided in practice and will impact upon the process of trading in much the same way as the latency buffer does.

Definition 2 (last look for on-rate trading) *A trader's request to sell at the indicative bid price published by LP-i at time t, and considered for execution by the LP at time $t+n$, is accepted at rate $b_t^{(i)}$ when:*

$$-\delta_- \leq b_{t+n}^{(i)} - b_t^{(i)} \leq \delta_+ \qquad \text{for } n \geq 0,\ \delta_- > 0,\ \delta_+ > 0, \tag{6}$$

and rejected otherwise. Analogously, a trader's request to buy at rate $a_t^{(i)}$ is accepted when $-\delta_+ \leq a_{t+n}^{(i)} - a_t^{(i)} \leq \delta_-$ and rejected otherwise.

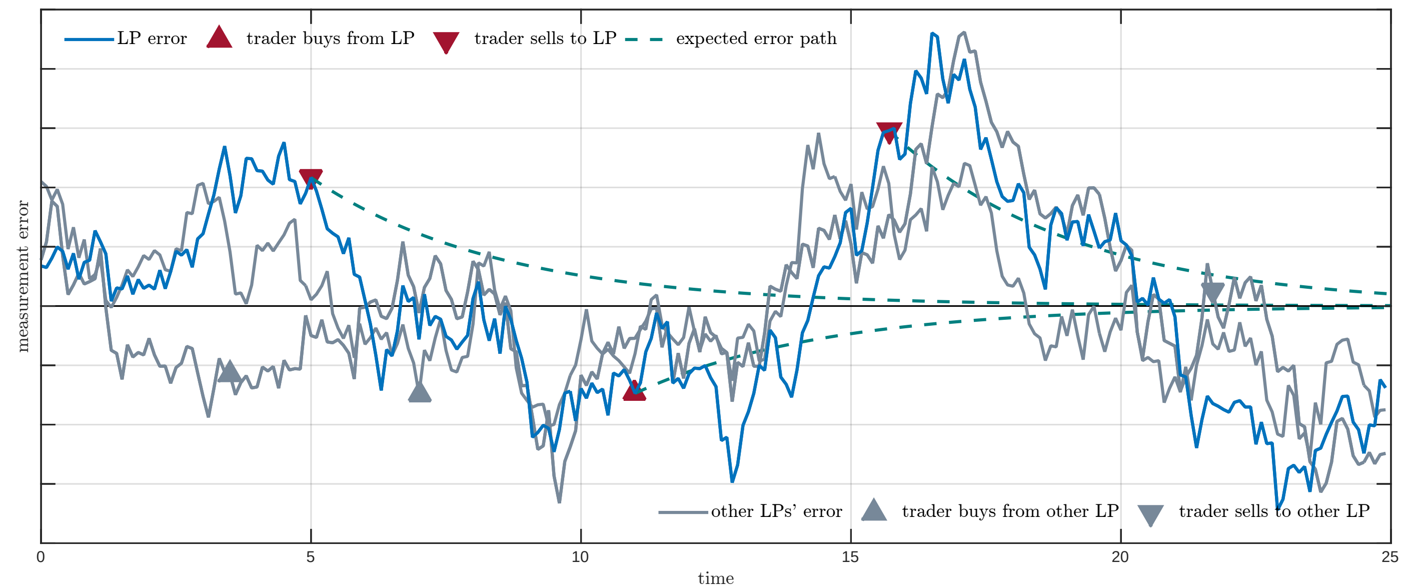
The threshold δ_- reflects the LP's tolerance to adverse price moves over the latency buffer while still accepting the deal request. So for sufficiently large δ_-, even when the trader is attempting to buy at what is at the time of decision a stale rate substantially below market, the LP will still accept the deal. The rationale for providing such a tolerance would be one where the trader is not systematically attempting to "pick off" the LP and the trades on stale quotes are as likely to go in the LP's favour as they are in the trader's favour. It is merely the randomness of the price process that determines this, rather than a deliberate latency motivated strategy on the trader's behalf. With latency sensitive flow, however, the LP may reduce the tolerance level δ_- and possibly increase the latency buffer n as a defensive measure.

The threshold δ_+ reflects the level up to which the LP accepts the deal request when the intervening price movement is in the LP's favour: once beyond that level the LP rejects the deal request and if the trader still desires liquidity she may re-attempt to deal at a rate that is likely – but not guaranteed – to be more favourable. So δ_+ provides the trader with a protection to prevent it from trading at unfavourable stale rates, but it comes at the cost of increased execution risk.

I refer to "no last look" when $\delta_\pm = \infty$, "one-sided last look" when only δ_- is set (and $\delta_+ = \infty$), and to "two-sided last look" when both δ_- and δ_+ are set to finite values. Figure 1 provides an illustration. It draws the distribution of the trade sign adjusted price move over the latency buffer, $\alpha_{t,n}^{(i)}$, following a deal request from a trader that adopts best price routing as in Definition 1.

$$\alpha_{t,n}^{(i)} \equiv \begin{cases} p_{t+n}^{(i)} - p_t^{(i)} & \text{for sell requests on } b_t^{(i)} > b_t^{(\neq i)} \\ -(p_{t+n}^{(i)} - p_t^{(i)}) & \text{for buy requests on } a_t^{(i)} < a_t^{(\neq i)} \end{cases} . \tag{7}$$

Figure 2: Illustration of best price routing and adverse selection in an aggregator



Note. This figure draws a sample path of the measurement error process $m_t^{(i)}$ for $N = 3$ homogeneous LPs and $\omega = 0.35, \beta = 0.75, \rho = 0.5$. Superimposed, is a set of deal requests from the uninformed trader routed to the LP with the best price (i.e. largest error in the desired direction). The dashed lines draw the expected future evolution of the measurement error from the points where deal requests are received.

When the trader deals exclusively with a single LP, then due to the uninformed nature of its flow, the distribution of α will be centred around zero. But when more than one LP is put in competition in an aggregator and the trader directs that same flow via best price routing, the distribution of α shifts to the left reflecting the fact that price moves subsequent to winning a deal request are more likely to go against the LP than in its favour. The LP is adversely selected: it does not receive a "random" set of deal requests from the trader, but only those when its price is more aggressive than any of the other competing LPs. It is highly probable that it has temporarily mis-priced the deal, and hence a subsequent corrective price adjustment is expected. An illustration of this is provided in Figure 2 and it explains why the mode of the distribution in Figure 1 is away from zero. See Oomen (2016) for a further discussion of these dynamics.

To analyse execution risk in this setup, I define the reject rate $\mathbb{R}$ as the probability of a deal request submitted by the trader following the best price routing rule in Definition 1 getting rejected by the originating LP on the basis of a specified last look rule.

$$\mathbb{R} \equiv \Pr(\text{LP-}i \text{ rejects deal request} \mid \text{LP-}i \text{ wins deal request}) \tag{8}$$

Note that because I assume the LPs are homogenous, the reject rate does not depend on i. From Eq. (8) it follows that the reject rate associated with the last look rule in Definition 2 is equal to:

$$\begin{aligned}
\mathbb{R} &= \Pr(b_{t+n}^{(i)} - b_t^{(i)} < -\delta_- \cup b_{t+n}^{(i)} - b_t^{(i)} > \delta_+ \mid b_t^{(i)} > b_t^{(\neq i)}) \\
&= \Pr(a_{t+n}^{(i)} - a_t^{(i)} < -\delta_+ \cup a_{t+n}^{(i)} - a_t^{(i)} > \delta_- \mid a_t^{(i)} < a_t^{(\neq i)})
\end{aligned}$$

Proposition 1 *For a panel of N homogenous liquidity providers competing for a trader's uninformed deal requests submitted at the best price as in Definition 1, and a trade acceptance rule as in Definition 2, the probability of a deal request getting rejected is approximately:*

$$\mathbb{R} = \mathbb{R}_- + \mathbb{R}_+ \qquad \textit{where} \qquad \mathbb{R}_- \approx \Phi\left(\frac{\mu_\alpha - \delta_-}{\sigma_\alpha}\right) \qquad \textit{and} \qquad \mathbb{R}_+ \approx 1 - \Phi\left(\frac{\mu_\alpha + \delta_+}{\sigma_\alpha}\right), \tag{9}$$

where $\mu_\alpha = (1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N$, $\sigma_\alpha^2 = n\sigma^2 + (1-\beta^{2n})\omega^2$, and $\Phi(\cdot)$ denotes the distribution of a standard normal random variable.

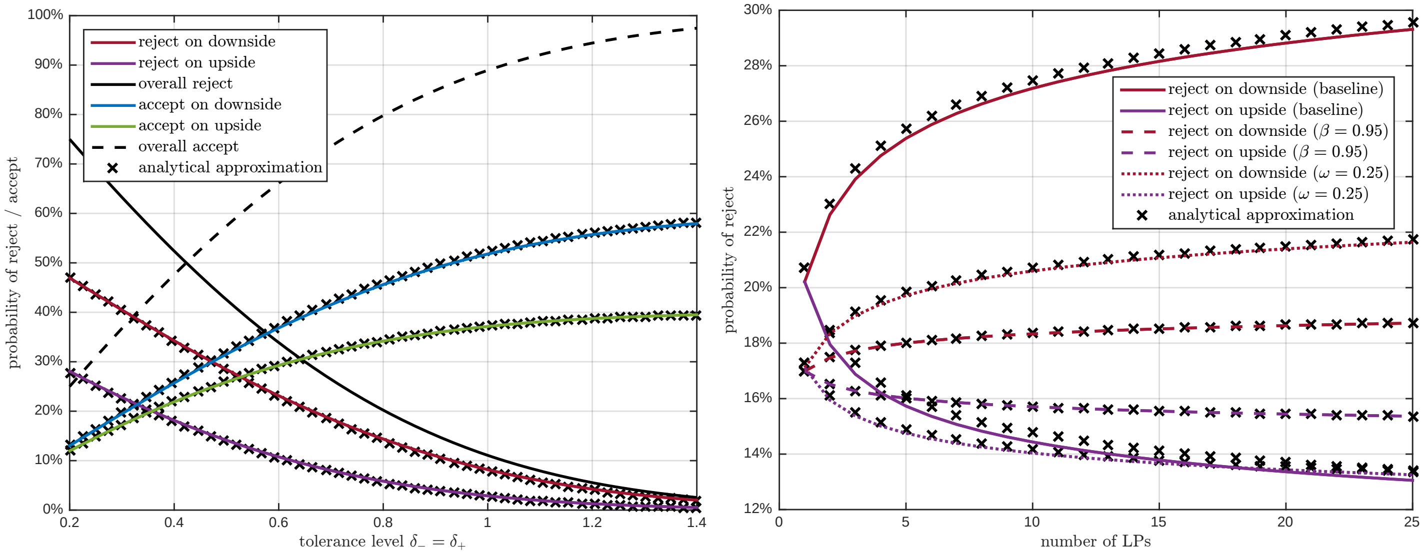
Proof Follows directly from Oomen (2016). ■

In Eq. (9), the term $\mathbb{R}_-$ represents the reject rate due to price movements that are adverse to the LP (i.e. the ■ left-tail in Figure 1) and $\mathbb{R}_+$ is the reject rate due to price movements that are favourable to the LP (i.e. the ■ right-tail in Figure 1). Clearly, the two-sided last look rule will always exhibit a higher reject rate than the one-sided last look rule provided that both set the same δ_- threshold. Similarly, when $\delta_- = \delta_+$ then $R_- > R_+$ because $\mu_\alpha > 0$. This is a consequence of the adverse selection induced by aggregation which shifts the mode of the price distribution over the latency buffer to the left and makes moves against the LP – and therefore rejects in the left-tail – more likely than moves in the LP's favour or rejects in the right-tail.

To illustrate, assume an aggregator setup with $N = 15$ homogenous LPs. The measurement error volatility, persistence, and correlation is set to $\omega = 0.5$, $\beta = 0.75$, $\rho = 0.5$ respectively, the nominal spread $s = 1$, and the latency buffer $n = 1$. The true price variation is assumed to be of same magnitude as the measurement error, i.e. $\sigma = 0.5$. In this case, the volatility of the trade-sign adjusted price move over the latency buffer is $\sigma_\alpha = 0.60$ with a mean of $-\mu_\alpha = -0.15$. This implies that the price is about 50% more likely to move over the latency buffer against the LP that wins the deal request than it is to move in its favour (i.e. 60% in trader's favour and 40% in LP's favour). If, as a defensive measure, it sets $\delta_- = \frac{1}{2}$ (i.e. it will reject a deal request when more than the entire intended spread capture is eroded over the last buffer) then it will reject 28.5% of deal requests on that side and if it also sets $\delta_+ = \frac{1}{2}$ in a two-sided setup then another 14.1% of deals will be rejected in the right-tail. When increasing the tolerance level to $\delta_- = 1$ and $\delta = 1$, these figures reduce to 8.1% and 2.9% respectively. See Figure 3 for further illustrations.

Figure 3: Reject rate for one-sided and two-sided last look

Panel A : reject and accept rate vs tolerance level   Panel B : reject rate vs number of LPs



Note. Panel A draws the reject rate as a function of last look tolerance levels (δ_- and δ_+) distinguishing between the situations where the price move over the latency buffer is adverse to the LP (i.e. "reject on downside" $\mathbb{R}_-$) and favourable to the LP (i.e. "reject on upside" $\mathbb{R}_+$). The corresponding accept rates are also included. The model parameters are set as $s = 1, \sigma = 0.5, \omega = 0.5, \beta = 0.75, \rho = 0.5, n = 1, N = 15$ and tolerance levels are varied between 0.2 and 1.4. Panel B draws the reject rates as a function of number of LPs (N) participating in the aggregator. The baseline scenario sets $\delta_- = \delta_+ = 0.5$ and the remaining model parameters as in Panel A. In the alternative scenarios β is increased and ω decreased.

One-sided last look is sometimes referred to as "asymmetric last look" and two-sided last look as "symmetric last look" (see, e.g., Bank of England, H.M. Treasury, and Financial Conduct Authority, 2015; Norges Bank Investment Management, 2015). I argue below that this terminology is ambiguous because several conflicting interpretations of symmetry exist and consequently many two-sided last look configurations may be considered asymmetric whilst certain one-sided last look designs may in fact be considered symmetric. The question of how symmetry can and should be defined is an interesting one, and within the model setup of this paper I will outline four last look designs that each can be viewed as symmetric according to specific well-defined criteria.

Symmetry in tolerance level. The simplest interpretation of last look symmetry is one that requires the upside and downside tolerance levels to be equal in value, i.e.

$$\delta_+ = \delta_-. \tag{10}$$

As discussed above, and made explicit in Proposition 1, symmetry in tolerance levels implies asymmetry in reject rates with $R_{-} > R_{+}$. This is due to the adverse selection induced by aggregation where moves in the trader's favour are more likely than those in the LP's favour and so with equal thresholds, rejects on the downside are more likely than those on the upside. See Panel A of Figure 4 for an illustration (as well as Figure 3).

Symmetry in reject probability. The above observation motivates an alternative way to define symmetry, i.e. set

$$\{\delta_{-}, \delta_{+}\} \quad \text{s.t.} \quad R_{-} = R_{+}. \tag{11}$$

That is, the tolerance levels are set in order to achieve balanced reject rates on trades that go in the LPs favour and those that go in the trader's favour. Using the analytical expression of the reject rate in Proposition 1, the requirement that $R_{-} = R_{+}$ is satisfied when $\delta_{+} = \delta_{-} - 2\mu_{\alpha}$. Because $\mu_{\alpha} > 0$, to attain symmetry in reject rates, asymmetry in tolerance levels is required with $\delta_{+} < \delta_{-}$. Panel B of Figure 4 illustrates this scenario.

Symmetry in incurred costs & revenues. If the price move over the latency buffer is within the tolerance levels set, the LP will accept the trade on the originally shown indicative rate even if the price has moved away. The LP (and trader) therefore incurs a marked-to-market cost when the price move is adverse and a revenue when the price move is favourable. This suggests another interpretation of symmetry where tolerance levels are set to balance the incurred costs and revenues (for both the trader and the LP) associated with price movements over the latency buffer. Formally, this requires one to set

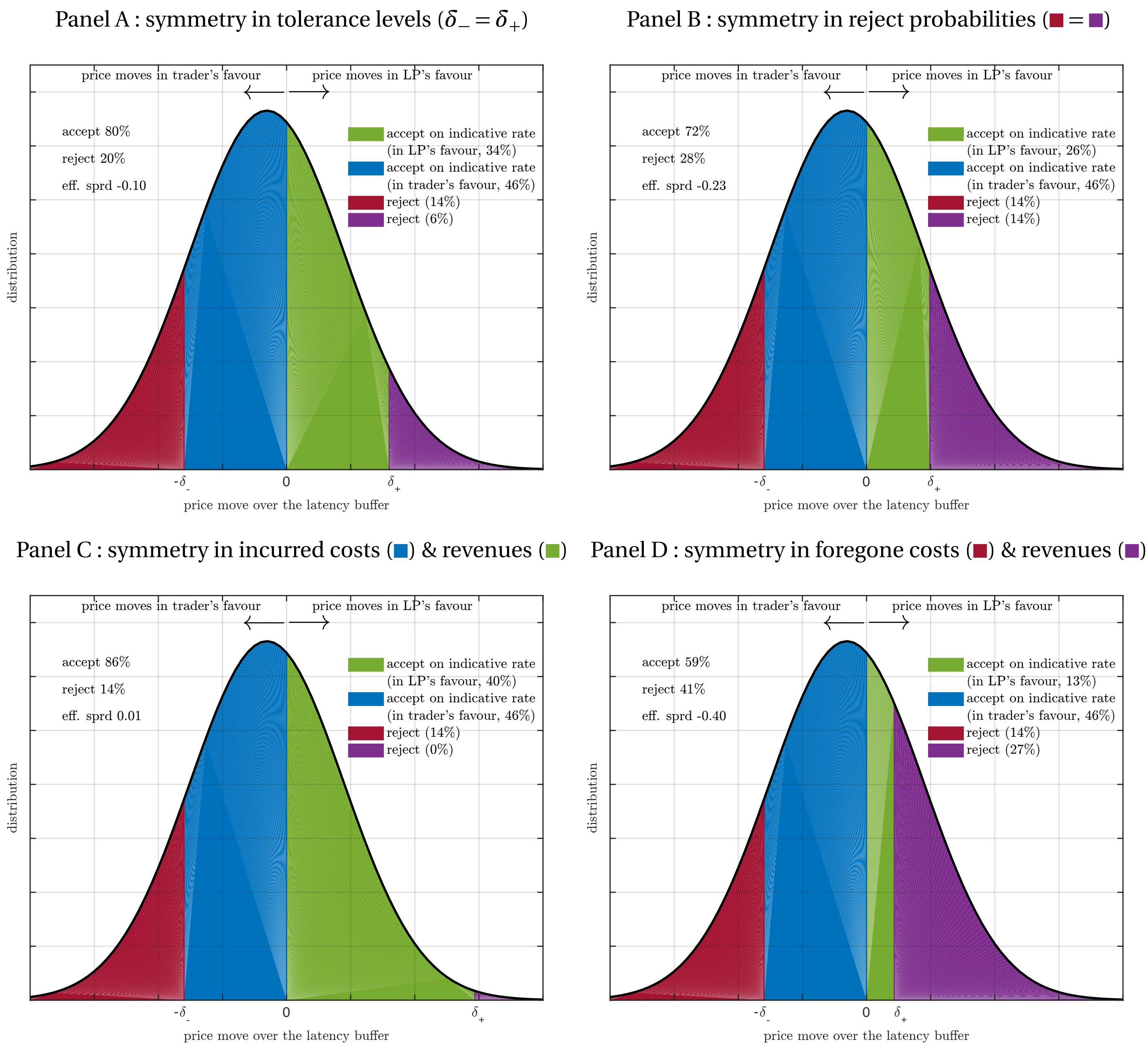
$$\{\delta_{-}, \delta_{+}\} \quad \text{s.t.} \quad -\int_{-\delta_{-}}^{0} \alpha f(\alpha) d\alpha = \int_{0}^{\delta_{+}} \alpha f(\alpha) d\alpha, \tag{12}$$

where α is as defined in Eq. (7) (suppressing dependence on t, n and i for notational convenience), and $f(\cdot)$ is the probability density function of α. Intuitively, Eq. (12) requires the tolerance levels to be set such that the price movement over the latency buffer integrated over the ■ area in Panel C of Figure 4 where the price moves in the trader's favour is equal and opposite to the integral taken over the ■ area where the price moves in the LP's favour.

Symmetry in foregone costs & revenues. A deal request that breaches the tolerance level δ_{-} (δ_{+}) leads to a reject that protects the LP (trader) and lets it avoid an otherwise costly trade when marked-to-market at the point of last look decision. The fourth and final interpretation of symmetry discussed here sets the tolerance levels such that the foregone costs associated with rejects in the left tail balance out the foregone revenues of rejects in the right tail, i.e. set

$$\{\delta_{-}, \delta_{+}\} \quad \text{s.t.} \quad -\int_{-\infty}^{-\delta_{-}} \alpha f(\alpha) d\alpha = \int_{\delta_{+}}^{\infty} \alpha f(\alpha) d\alpha. \tag{13}$$

Figure 4: Illustration of different symmetrical last look designs



Note. This figure draws the distribution of price moves over the latency buffer as in Figure 1 for different interpretations of symmetry when setting the tolerance levels for the two-sided last look on-rate trading protocol. The model parameters are set to $s = 1, \sigma = 0.5, \omega = 0.5, \beta = 0.75, \rho = 0.5, n = 1, N = 15$. Panel A enforces symmetry in tolerance levels, i.e. $\delta_+ = \delta_-$. Panel B sets the tolerance levels as in Eq. (11) to ensure that the probability of a reject on the downside ($\mathbb{R}_-$) is equally likely as one on the upside ($\mathbb{R}_+$). This is achieved when $\delta_+ \approx 0.62\delta_-$ for the current parametrisation of the model. In Panels C and D the last look tolerance levels are set according to Eq. (12) and (13) in order to balance the incurred and foregone costs associated with price moves against the LP with incurred and foregone revenues associated with favourable price moves. This is achieved when $\delta_+ \approx 1.83\delta_-$ and $\delta_+ \approx 0.27\delta_-$ respectively. The effective spread defined in Section 4 is also reported for each scenario.

Panel D of Figure 4 provides an illustration where the foregone costs in the ■ area are balanced out by the foregone revenues in the ■ area.

To conclude, Figure 5 draws the tolerance levels for these four interpretations of symmetry[6] in Panel A and the associated reject rates in Panel B. It is important to note that symmetry cannot always be attained. For instance, for sufficiently small δ_-, symmetry in reject probability cannot be achieved as it would require $\delta_+ < 0$. The same limitation is true for symmetry in foregone costs and revenues. Conversely, enforcing symmetry in incurred costs and revenues implicitly sets a ceiling on the tolerance level δ_- that can be granted to the trader: the distribution of price moves over the latency buffer is centred to the left which means that for certain δ_- there simply is not enough probability mass on the right to generate marked-to-market revenues that make up for the costs incurred on the left. This leaves symmetry in tolerance levels – despite it being distinctly asymmetric in terms of reject rates, foregone and incurred revenues and costs – as the only interpretation that is universally applicable. From an implementation perspective it is also the only one that does not require knowledge of the price distribution over the latency buffer (which can be estimated but is not known in practice) to set the tolerance levels. As a final remark, when the distribution of α is centred on zero (i.e. in the model this occurs for $N = 1$ when the trader deals with a single LP exclusively) symmetry by all four criteria is achieved by setting $\delta_- = \delta_+$. But when the distribution is not centred on zero, then the different versions of symmetry are mutually exclusive so that whichever way one sets the tolerance levels, it'll be asymmetric by at least three criteria.

3.2 Price improvements

An alternative to the last look rule in Definition 2, which may reject deal requests when the price moves in the trader's favour, is to instead accept these requests but at a revised rate that incorporates a rebate to the trader for the incremental price improvement beyond δ_+ observed over the latency buffer.

Definition 3 (last look for on-rate trading with rebate) *A trader's request to sell at the indicative bid price published by LP-i at time t, and considered for execution by the LP at time t + n, is accepted at rate $b_t^{(i)}$ when:*

$$-\delta_- \leq b_{t+n}^{(i)} - b_t^{(i)} \leq \delta_+ \qquad \text{for } n \geq 0,\ \delta_- > 0,\ \delta_+ > 0, \tag{14}$$

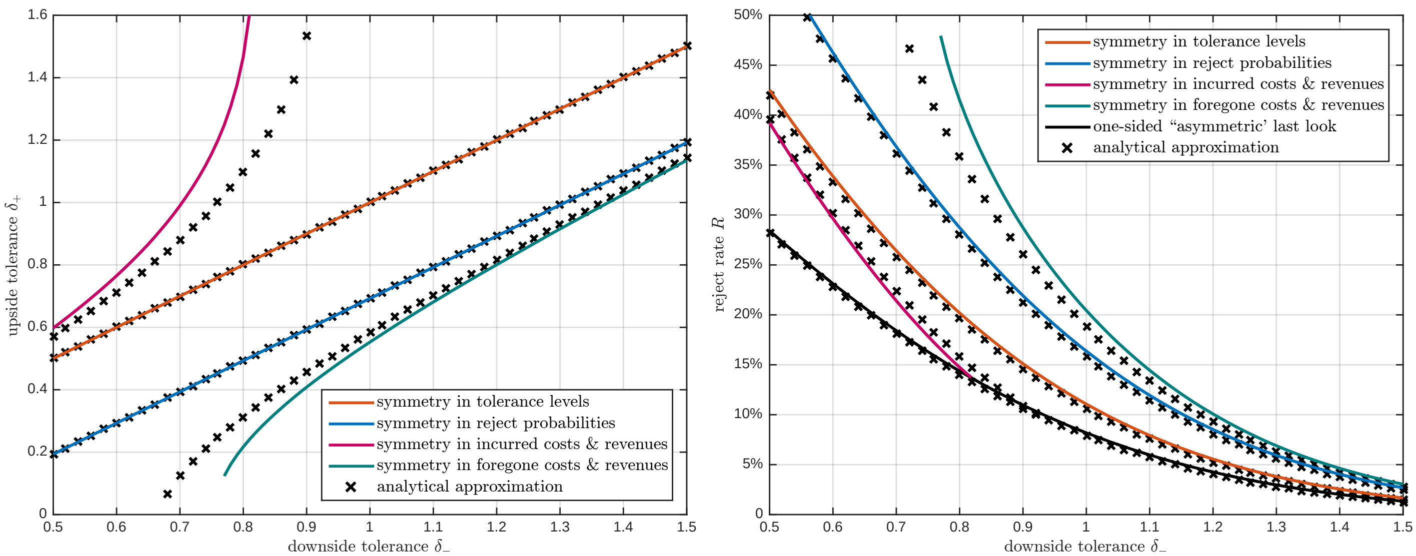
accepted at rate $b_t^{(i)}$ plus a rebate of $b_{t+n}^{(i)} - b_t^{(i)} - \delta_+$ if:

$$b_{t+n}^{(i)} - b_t^{(i)} > \delta_+, \tag{15}$$

[6] Using the results presented in Section 4, for a given threshold δ_-, an analytical approximation for δ_+ can be obtained by numerically solving $G(0,\delta_+,\sigma_\alpha^2)\left(\Phi\left(\frac{\mu_\alpha+\delta_+}{\sigma_\alpha}\right)-\Phi\left(\frac{\mu_\alpha}{\sigma_\alpha}\right)\right)=-G(-\delta_-,0,\sigma_\alpha^2)\left(\Phi\left(\frac{\mu_\alpha}{\sigma_\alpha}\right)-\Phi\left(\frac{\mu_\alpha-\delta_-}{\sigma_\alpha}\right)\right)$ when symmetry in incurred costs and revenues is desired whereas symmetry in foregone costs and revenues is obtained by setting δ_+ to solve $G(\delta_+,\infty,\sigma_\alpha^2)\left(1-\Phi\left(\frac{\mu_\alpha+\delta_+}{\sigma_\alpha}\right)\right)=-G(-\infty,-\delta_-,\sigma_\alpha^2)\Phi\left(\frac{\mu_\alpha-\delta_-}{\sigma_\alpha}\right)$.

Figure 5: Illustration of tolerance levels and reject rates for different symmetrical last look designs

Panel A : symmetrical tolerance levels (δ_+ vs δ_-) Panel B : reject rates under symmetry



Note. Panel A draws the upside last look tolerance level δ_+ as a function of δ_- for the four different interpretations of symmetry as in Eqn. (10–13). Panel B draws the reject rate for the different symmetrical two-sided last look designs as a function of δ_- (with δ_+ set according to Panel A). The model parameters are set to $s = 1, \sigma = 0.5, \omega = 0.5, \beta = 0.75, \rho = 0.5, n = 1, N = 15$. In both panels, the solid lines are the true values obtained by simulation and the black crosses indicate the analytical approximations.

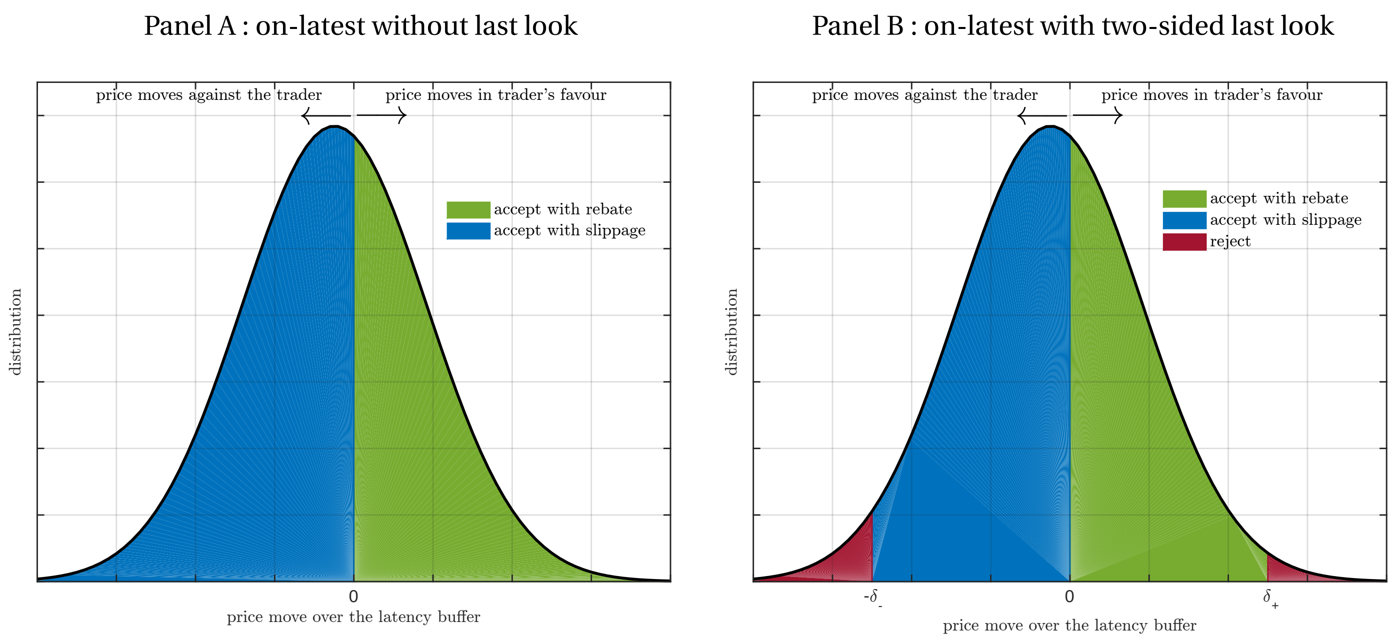
and rejected otherwise. Analogously, a trader's request to buy is accepted when $-\delta_+ \leq a_{t+n}^{(i)} - a_t^{(i)} \leq \delta_-$*, accepted with a rebate of* $a_{t+n}^{(i)} - a_t^{(i)} - \delta_+$ *if* $a_{t+n}^{(i)} - a_t^{(i)} < -\delta_+$*, and rejected otherwise.*

An appealing feature of this approach is that the probability of a reject is now reduced to R_- whereas the probability of getting a deal request accepted with a rebate is R_+ (i.e. the ■ area in Panel B of Figure 1) . Certainty of execution is reduced in exchange for some uncertainty of rate. Note that, assuming the deal request is accepted, the trader now sells at $\max(b_t^{(i)}, b_{t+n}^{(i)} - \delta_+)$ and buys at $\min(a_t^{(i)}, a_{t+n}^{(i)} - \delta_+)$. In line with the discussion above, numerous interpretations of symmetry can be entertained for this last look design.

3.3 Trading on the latest rate

As previously mentioned, on-rate trading provides certainty of rate but lacks certainty of execution. Some traders may value certainty of execution over certainty of rate, and this is what on-latest trading offers.

Figure 6: Last look for on-latest trading



Note. This figure draws the distribution of the trade sign adjusted price move (i.e. $b^{(i)}_{t+n} - b^{(i)}_t$ for requests to buy and $-(a^{(i)}_{t+n} - a^{(i)}_t)$ for requests to sell) over the latency buffer n immediately following a deal request from a trader that adopts best price routing as in Definition 1. The shared areas indicate the regions where for the on-latest trading protocol as in Definition 4, a rebate (■) or slippage (■) is passed on following a favourable or adverse price move from the trader's perspective. In Panel B, two-sided last look is applied and requests are rejected on price moves beyond the set tolerance levels in the ■ tail-areas.

Definition 4 (last look for on-latest trading) *A trader's request to sell on the basis of – but not insisting on – the indicative bid price published by LP-i at time t, and considered for execution by the LP at time $t+n$, is accepted at rate $b^{(i)}_{t+n}$ when:*

$$-\delta_- \leq b^{(i)}_{t+n} - b^{(i)}_t \leq \delta_+ \qquad \textit{for } n \geq 0,\ \delta_- > 0,\ \delta_+ > 0, \tag{16}$$

and rejected otherwise. Analogously, a trader's request to buy is accepted at rate $a^{(i)}_{t+n}$ when $-\delta_+ \leq a^{(i)}_{t+n} - a^{(i)}_t \leq \delta_-$ and rejected otherwise.

The simplest setup is one where $\delta_- = \delta_+ = \infty$ as in Panel A of Figure 6. Here, all deal requests are accepted and so any uncertainty of execution is eliminated. It can be argued that this protocol offers "symmetric" last look in that any slippage and price improvement is passed on to the trader in full. In contrast to on-rate trading, the LP is not conflicted in that it is indifferent between the price moving against or in favour of the trader over the latency buffer interval.

In practice, the trader may want to protect herself against dealing at prices that are excessively far away from where the request was submitted. For instance, it is unlikely that the trader's liquidity demand is independent of the price level: a request to sell submitted on the basis of a bid at $b_t^{(i)}$ is unlikely to still be desired in the same quantity when it gets filled at $b_{t+n}^{(i)} << b_t^{(i)}$ (e.g. trader may have wanted to hold out for a better price instead) or at $b_{t+n}^{(i)} >> b_t^{(i)}$ (e.g. in the case of a take-profit order, the trader may have preferred to not close out the position). Such scenarios can occur in fast moving markets over news announcements or market dislocations (e.g. US non-farm payrolls, the 2010 equity market flash crash, the January 2015 SNB discontinuation of the Swiss Franc cap). Similar situations may arise when technology problems prevent the trader from observing up-to-date price information and any resulting deal requests may be submitted at severely off-market rates. To address this, last look tolerance levels $\delta_\pm$ can be set in much the same way as before although the motivation is somewhat different: the thresholds are primarily intended to serve as a circuit breaker control more so than to act as a defensive tool for the LP to manage the adverse selection or protect against latency sensitive flow. Because the underlying price dynamics are unchanged, the probability of reject for on-latest with two-sided last look is still characterised by Proposition 1 above.

4 Transaction cost analysis

The various trading protocols and last look designs discussed in the preceding section span a continuum of execution risk profiles as characterised by the degree of execution and rate uncertainty faced by the trader. Table 1 provides a summary. The question that naturally arises now is which execution setup is "best"? This clearly depends on the subjective and trader specific preferences and in that sense there is no simple answer. But what can be analysed are measures of transaction costs paid by the trader under the different scenarios. A natural quantity to consider in this context is the effective spread proposed by Oomen (2016). It is defined as (twice) the expected distance the traded price is away from the true mid price at the point of execution, i.e.

$$\begin{aligned} \mathbb{S} &\equiv 2E(\text{true mid at point of execution} - \text{transacted bid price} \mid \text{win deal request \& accept deal request}), \\ &= 2E(\text{transacted offer price} - \text{true mid at point of execution} \mid \text{win deal request \& accept deal request}). \end{aligned} \tag{17}$$

So for the on-rate trading protocol, the effective spread is given by:

$$\begin{aligned} \mathbb{S} &= 2E\left(p^*_{t+n} - b_t^{(i)} \mid \text{LP wins deal request \& accepts deal request}\right), \\ &= \lim_{h\to\infty} 2E\left(p^{(i)}_{t+h} - b_t^{(i)} \mid \text{LP wins deal request \& accepts deal request}\right). \end{aligned} \tag{18}$$

Table 1: Execution risk for different trading protocols and last look designs

		certainty of rate	
		Yes	No
certainty of execution	Yes	on-rate with $\delta_\pm = \infty$ i.e. no last look	on-latest with $\delta_\pm = \infty$ and on-rate with rebate and $\delta_- = \infty$
	No	on-rate with finite δ_- or δ_+	on-latest with finite δ_- or δ_+ and on-rate with rebate and finite δ_-

Note. This table summarises the trading protocols and last look designs as in Definitions 2, 3, and 4, categorised by the degree of execution risk in terms of certainty of rate and certainty of execution.

While the true price process p^* is unobserved, Eq. (18) provides a feasible estimator. For the on-latest protocol, the effective spread is defined as:

$$\mathbb{S} = 2E\left(p^*_{t+n} - b^{(i)}_{t+n} \mid \text{LP wins deal request \& accepts deal request}\right). \tag{19}$$

Because $\mathbb{S}$ represents the effective spread the LP earns from the trade, by reflection it represents the effective spread the trader pays.

Theorem 1 *For a panel of N homogenous liquidity providers competing for a trader's uninformed deal requests submitted at the best price as in Definition 1, and a trading protocol as in Definition 2 (i.e. last-look for on-rate trading), an approximation of the effective spread is:*

$$\mathbb{S} \approx s - 2\omega\sqrt{1-\rho}\,\psi_N + 2\frac{n\sigma^2}{\sigma^2_\alpha} G(\mu_\alpha - \delta_-, \mu_\alpha + \delta_+, \sigma^2_\alpha), \tag{20}$$

where

$$G(l, u, \sigma^2) = \sigma \frac{\phi(l/\sigma) - \phi(u/\sigma)}{\Phi(u/\sigma) - \Phi(l/\sigma)}, \tag{21}$$

and $\mu_\alpha = (1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N$, $\sigma^2_\alpha = n\sigma^2 + (1-\beta^{2n})\omega^2$, $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution of a standard normal random variable. When $\delta_+ = \infty$, the expression in Eq. (20) constitutes a lower-bound on the effective spread. For a trading protocol as in Definition 3 (i.e. last-look for on-rate trading with rebate), an approximation of the effective spread is:

$$\mathbb{S} \approx s - 2\omega\sqrt{1-\rho}\,\psi_N + 2\frac{n\sigma^2}{\sigma^2_\alpha} G(\mu_\alpha - \delta_-, \infty, \sigma^2_\alpha) - 2\left(G(\mu_\alpha + \delta_+, \infty, \sigma^2_\alpha) - (\mu_\alpha + \delta_+)\right)\frac{\mathbb{R}_+}{1-\mathbb{R}_-}. \tag{22}$$

Finally, for a trading protocol as in Definition 4 (i.e. last look for on-latest trading), an approximation of the effective spread is:

$$\mathbb{S} \approx s - 2\beta^n \omega\sqrt{1-\rho}\psi_N - 2(1-\beta^{2n})\frac{\omega^2}{\sigma_\alpha^2} G(\mu_\alpha - \delta_-, \mu_\alpha + \delta_+, \sigma_\alpha^2). \tag{23}$$

When $\delta_- = \delta_+ = \infty$, then $G(\cdot) = 0$ and the expression in Eq. (23) is an exact expression of the effective spread.

Proof See Appendix A. ∎

The above theorem makes explicit the different components that make up the effective spread. For on-rate trading, Eq. (20) can be further decomposed as follows:

$$\mathbb{S} \approx \underbrace{s}_{\text{nominal spread (+)}} -2 \underbrace{\omega\sqrt{1-\rho}\psi_N}_{\text{adverse selection (-)}} + 2\frac{n\sigma^2\sigma_\alpha^{-2}}{1-\mathbb{R}_- - \mathbb{R}_+}(\underbrace{G(\delta_- - \mu_\alpha, \infty, \sigma_\alpha^2)}_{\text{last look protection (+)}}\mathbb{R}_- + \underbrace{G(-\infty, -\mu_\alpha - \delta_+, \sigma_\alpha^2)}_{\text{foregone revenues (-)}}\mathbb{R}_+).$$

This highlights four distinct contributors to the effective spread, namely (i) the nominal spread, (ii) the spread compression due to aggregation or equivalently the strength of adverse selection, and (iii) recovered adverse selection costs via the one-sided last look process, and (iv) foregone revenues on the rejects due to the δ_+ tolerance level. Note that for one-sided last look, this last component is zero whereas for two-sided last look and the fourth interpretation of symmetry, the third and fourth component balance out. For on-rate trading with rebates in Eq. (22), another component is added that measures the cost to the LP of foregoing some revenues and paying it in the form of a rebate to the trader. Panel A of Figure 7 provides an illustration of the effective spread for one-sided last look and two-sided last look with and without a rebate, together with its components.

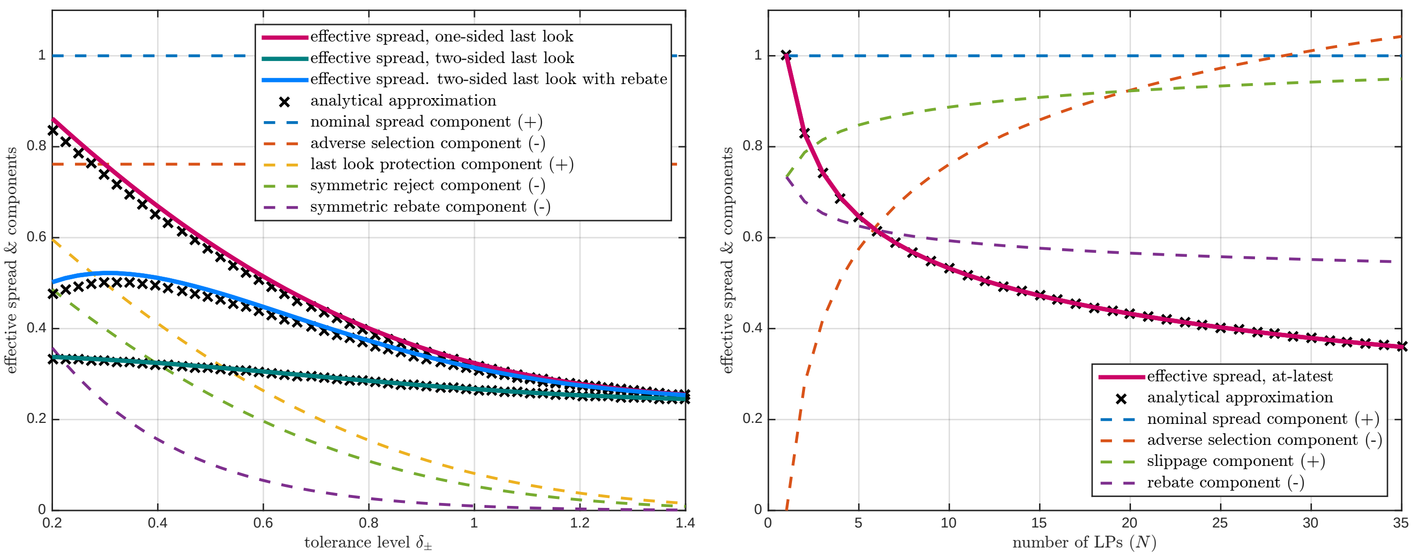
For on-latest trading, first consider the no last look case where $\delta_\pm = \infty$ (note the slight abuse of terminology here, because for at-latest, even when $\delta_\pm = \infty$, the last look process is still required to determine the rate). The effective spread in Eq. (23) can be expressed and decomposed as follows:

$$\begin{aligned} \mathbb{S} &= s - 2\beta^n\omega\sqrt{1-\rho}\psi_N, \\ &= s - 2\omega\sqrt{1-\rho}\psi_N + 2\underbrace{(\mu_\alpha - G(-\infty, \mu_\alpha, \sigma_\alpha^2))}_{\approx \text{ expected slippage}} \underbrace{\Phi(\mu_\alpha/\sigma_\alpha)}_{\approx \text{Pr. of slippage}} - 2\underbrace{(G(\mu_\alpha, \infty, \sigma_\alpha^2) - \mu_\alpha)}_{\approx \text{ expected rebate}} \underbrace{(1-\Phi(\mu_\alpha/\sigma_\alpha))}_{\approx \text{Pr. of rebate}}. \end{aligned} \tag{24}$$

Because the distribution of trade-sign adjusted price moves over the latency buffer is shifted towards the left, the trader's marked-to-market costs associated with incurring slippage exceed the revenues earned from price improvements. This is illustrated in Panel B of Figure 7. For the on-rate trading protocol, the last look tolerance levels $\delta_\pm$ are a key lever to guard against excessive adverse selection. This is different for the on-latest trading protocol where $\delta_\pm$ play a less important role and the LP protection primarily comes from the trading protocol itself and the choice

Figure 7: Illustration of the effective spread and its components for on-rate and on-latest trading

Panel A : effective spread vs $\delta_{\pm}$ for on-rate trading   Panel B : effective spread vs N for on-latest trading



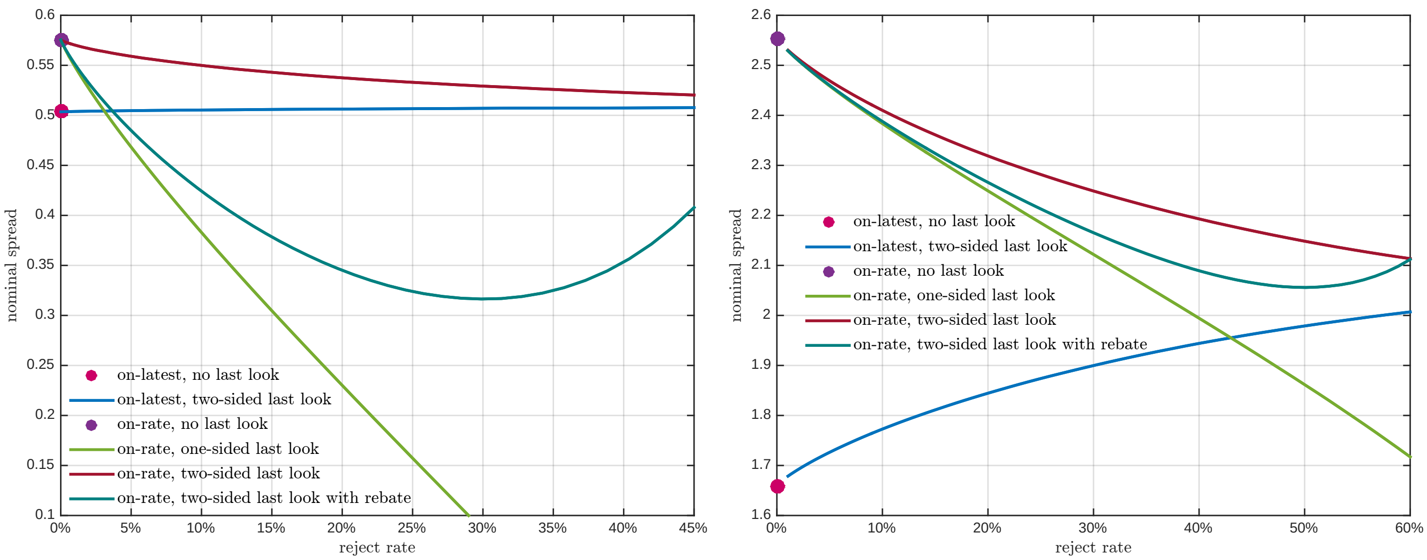
Note. Panel A draws the effective spread for on-rate trading as a function of tolerance levels $\delta_{\pm}$ for one- and two-sided last look as in Definition 2, and two-sided last look with rebate as in Definition 3, together with its components as detailed in Theorem 1 and the text. Panel B draws the effective spread and its components for on-latest trading without last look as a function of the number of LPs. The model parameters are set as $s = 1, \sigma = 0.5, \omega = 0.35, \beta = 0.85, \rho = 0.5$. Additionally, in Panel A, $n = 1, N = 10$ and $\delta_{-} = \delta_{+}$ is varied between 0.2 and 1.4 whereas in Panel B $n = 3, \delta_{\pm} = \infty$ and N is varied between 1 and 35.

of latency buffer n. Note that the adverse selection component in Eq. (23) is multiplied by β^n and so it will reduce exponentially with n.

To conclude, it is instructive to consider for each of the different trading protocols and last look methodologies, what combinations of nominal spread s and reject rate $\mathbb{R}$ (as determined by the tolerance levels $\delta_{\pm}$) result in the same effective spread. A trader that measures her execution performance solely by the effective spread would be indifferent between all the identified configurations. Figure 8 provides an illustration. It traces out the points in $(s, \mathbb{R})$ space that deliver for each trading protocol and last look design a break-even effective spread of $\mathbb{S} = 0$. Panel A considers a setup where – from the LPs' perspective – the trader's flow is relatively benign: she only puts a moderate number of LPs in competition for her flow ($N = 5$) whilst the LPs' ability to discover the true price is reasonably good ($\omega < \sigma$) and disagreement on price is limited ($\rho = 0.5$). In such a setting, the different last look methodologies can lead to materially different execution risk profiles despite attaining the same effective spread. For instance, assuming a reject rate of 20% for on-rate trading, an LP adopting one-sided last look needs to charge a nominal

Figure 8: Transaction cost indifference curves

Panel A : benign flow (moderate adverse selection) Panel B : adverse flow (strong adverse selection)



Note. This figure draws the set of trading protocol and last look configurations – as characterised by reject rate $\mathbb{R}$ and nominal spread s – that each incur identical transaction costs as measured by the effective spread $\mathbb{S}$ as defined in Theorem 1. In Panel A the model parameters are set to reflect a moderate amount of adverse selection in the trader's flow (i.e. $\sigma = 0.5, \omega = 0.35, \beta = 0.875, \rho = 0.5, n = 1, N = 5$). In Panel B, the model parameters are adjusted to simulate trader's flow that is experienced as highly adverse by the LP (i.e. $\sigma = 0.5, \omega = 0.75, \beta = 0.65, \rho = 0.25, n = 1, N = 25$). In all cases, the nominal spread s and last look tolerance levels $\delta_{\pm}$ are varied in order to attain an break-even effective spread $\mathbb{S} = 0$. For two-sided last look the tolerance levels are set to $\delta_- = \delta_+$.

spread of approximately 0.225 in order to achieve a zero effective spread, but an LP that uses two-sided last look will need to charge a nominal spread nearer to 0.55 and drops to 0.35 if rebates are permitted. Similarly, for a fixed nominal spread of 0.35, the LP that uses one-sided last look will reject 12.5% of deal requests while the LP that uses two-sided last look with rebates will reject 20% in order to achieve the same effective spread. For on-latest trading, note that a nominal spread of 0.5 results in a break-even effective spread almost regardless of the reject rate experienced as the last look tolerance levels are varied. The mathematical intuition is that the distribution of α is only marginally shifted to the left which means that (i) for on-latest the last look thresholds $\delta_- = \delta_+$ clip off about an equal mass in each tail of the distribution which affects the reject rate but does not materially affect the effective spread, and (ii) for on-rate trading the probability of breaching δ_- is about the same as breaching δ_+ and this makes two-sided versions of last look expensive relative to a one-sided implementation that does not need to rebate or forego revenues in that region.

What happens when the trader's flow turns materially adverse to the LP? The model assumes an uninformed trader, but by modifying the aggregator setup it is possible to generate flow that is observationally equivalent in that it is perceived as highly informed or directional by the LPs. Specifically, I increase the number of competing LPs ($N = 25$), reduce their ability to discover the true price ($\omega > \sigma$) and increase price disagreement amongst them ($\rho = 0.25$). To emphasise the aggressiveness of this setup, note that if each LP charges a nominal spread of $s = 1$, the trader will be observing nearly permanently inverted bid-offer prices in her aggregator at an average inside spread of $S = -1.6$. If the LPs set last look tolerance levels $\delta_- = \delta_+ = 1$, then rejects due to a adverse price movements will be nearly eight times more likely than those due to a favourable price movement ($R_- \approx 23.9\%$ and $R_+ \approx 3.1\%$) and on accepted trades the price will be twice as likely to go against the LP as in its favour. This is a setup where last look is a key defensive tool and protects the LP from what would otherwise be a guaranteed arbitrage. Panel B of Figure 8 traces out the combinations of reject rate and nominal spread that lead to the same effective spread of zero for the various trading protocols and last look designs. As expected, the nominal spread that is required by the LPs to break even is substantially higher than before: ranging between 1.65 and 2.55 depending on the setup. However, the key observation is that the distinction between the different last look designs fades and the primary driver of execution risk is reduced to the choice of trading protocol. Intuitively, in an highly aggregated setup (or similarly, when trading is driven by short-term price predictions, or aggressive stack-sweep execution is adopted) the adverse selection experienced by the LPs can be very strong and following a deal request the price is much more likely to go against the LP than it is in its favour. Providing a rebate or rejecting a deal in the rare instances where the price moves against the trader is therefore relatively inexpensive for the LP and consequently of limited value to the trader. In extremum, one-sided last look for on-rate trading will be observationally equivalent to any variant of two-sided last look.

For the on-latest protocol the situation is subtly different. Here, somewhat counterintuitively, the nominal spread the LP needs to charge to breakeven increases the more deal requests it rejects. The mechanics are as follows. With latency sensitive flow, the price will on average move strongly against the LP but do so in a stochastic (rather than a deterministic) manner. With very tightly set last look tolerance levels, the trader effectively sub-selects those sample paths where the price over the latency buffer remains approximately unchanged. That in itself does not remove the inherent adverse selection but merely ensures that the adverse price move is realised post latency buffer. At that point, however, the risk has been transferred and the LP will now bear the costs instead of being able to pass it on as slippage to the trader. To remain at breakeven, a wider normal spread is thus required.

5 Concluding remarks

Spot foreign exchange makes up the world's largest and most liquid financial market – about ten times the size of the US cash equity markets by notional volume – and yet it is traded in a non-centralised manner, on a disclosed and bi-lateral basis, as an over-the-counter product that allows for customisation. While currencies itself are naturally a standardised flow product, it is the provision of liquidity in that product that is often tailored to the individual trader. The typical setup is one where a panel of liquidity providers compete for a trader's order flow and offer bespoke liquidity specific to that trader, e.g. in terms of quoted spreads, available amounts, consistency of price, refresh frequency of quotes, delivery mechanism, ancillary services. Another important aspect of the trading process that can be negotiated – and the focus of this paper – is how the execution risk at a micro-structure level is shared between the liquidity provider and the trader. Last look and the choice of trading protocol are the mechanisms that controls this. I outline a framework in which to analyse different designs – including asymmetric and symmetric versions of last look and trading protocols that allow for slippage and rebates to be passed on to the trader – and provide a detailed comparative analysis in term of execution risk and effective transaction costs. The main finding is that there is no single design that is either superior or inferior, and that the choice of last look logic and trading protocol comes down to preferences over risk sharing.

A Proofs

Proof of Theorem 1. I first state some standard results on conditional expectations of normal random variables (see, e.g., Stuart and Ord, 1994) that will be used below. Let (x,y) be bi-variate normal with mean zero, variances σ_x^2, σ_y^2, and correlation ρ. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution function of a standard normal random variable. When $\sigma_x = 1$,

$$E(x \mid \delta_- < x < \delta_+) = \frac{E(X I_{(\delta_- < x < \delta_+)})}{E(I_{(\delta_- < x < \delta_+)})} = \frac{\int_{\delta_-}^{\delta_+} x\phi(x)dx}{\Phi(\delta_+) - \Phi(\delta_-)} = \frac{-\phi(x) \mid_{\delta_-}^{\delta_+}}{\Phi(\delta_+) - \Phi(\delta_-)} = \frac{\phi(\delta_-) - \phi(\delta_+)}{\Phi(\delta_+) - \Phi(\delta_-)}.$$

And so for arbitrary σ_x,

$$E(x \mid \delta_- < x < \delta_+) = \sigma_x E\left(x/\sigma_x \mid \delta_-/\sigma_x < x/\sigma_x < \delta_+/\sigma_x\right) = \sigma_x \frac{\phi(\delta_-/\sigma_x) - \phi(\delta_+/\sigma_x)}{\Phi(\delta_+/\sigma_x) - \Phi(\delta_-/\sigma_x)} \equiv G(\delta_-, \delta_+, \sigma_x^2), \tag{25}$$

where the function G is defined for notational convenience. Note that $G(-\infty, \infty, \sigma_x^2) = 0$ and $G(0, \infty, \sigma_x^2) = \sqrt{2/\pi}$. For a bi-variate normal,

$$E(x \mid \delta_- < y < \delta_+) = E\left(E_y(x \mid y) \mid \delta_- < y < \delta_+\right) = \frac{\rho\sigma_x\sigma_y}{\sigma_y^2} E(y \mid \delta_- < y < \delta_+) = \frac{\rho\sigma_x\sigma_y}{\sigma_y^2} G(\delta_-, \delta_+, \sigma_y^2). \tag{26}$$

Starting with the definition of the effective spread in Eq. (18) and applying the on-rate last look protocol as in Definition 2, gives:

$$\begin{aligned}
\mathbb{S} &= s + 2\lim_{h\to\infty} E\left(p_{t+h}^{(i)} - p_t^{(i)} \mid b_t^{(i)} > b_t^{(\neq i)}, -\delta_- < b_{t+n}^{(i)} - b_t^{(i)} < \delta_+\right), \\
&= s + 2\lim_{h\to\infty} E\left((\beta^h - 1)m_t^{(i)} + \sum_{j=0}^{h-1} \beta^j \eta_{t+h-j}^{(i)} + \sum_{j=1}^{h} \varepsilon_{t+j} \mid m_t^{(i)} > m_t^{(\neq i)}, -\delta_- < b_{t+n}^{(i)} - b_t^{(i)} < \delta_+\right).
\end{aligned} \tag{27}$$

An approximation for the conditional expectation of $m_t^{(i)}$ is:

$$E\left(m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}, -\delta_- < b_{t+n}^{(i)} - b_t^{(i)} < \delta_+\right) \approx E\left(m_t^{(i)} \mid m_t^{(i)} > m_t^{(\neq i)}\right) = \omega\sqrt{1-\rho}\,\psi_N. \tag{28}$$

Next, a lower bound on the second and third term in Eq. (27) for $h \geq n$ is given by:

$$\begin{aligned}
&E\left(E\left(\sum_{j=1}^{h} \varepsilon_{t+j} + \sum_{j=0}^{h-1} \beta^j \eta_{t+h-j}^{(i)} \mid -\delta_- < (\beta^n - 1)m_t^{(i)} + \sum_{j=1}^{n} \varepsilon_{t+j} + \sum_{j=0}^{n-1} \beta^j \eta_{t+n-j}^{(i)} < \delta_+\right) \mid m_t^{(i)} > m_t^{(\neq i)}\right) \\
&= E\left(E\left(\sum_{j=1}^{n} \varepsilon_{t+j} + \beta^{h-n} \sum_{j=0}^{n-1} \beta^j \eta_{t+n-j}^{(i)} \mid (1-\beta^n)m_t^{(i)} - \delta_- < \sum_{j=1}^{n} \varepsilon_{t+j} + \sum_{j=0}^{n-1} \beta^j \eta_{t+n-j}^{(i)} < (1-\beta^n)m_t^{(i)} + \delta_+\right) \mid m_t^{(i)} > m_t^{(\neq i)}\right), \\
&= \frac{n\sigma^2 + \beta^{h-n}(1-\beta^{2n})\omega^2}{n\sigma^2 + (1-\beta^{2n})\omega^2} E\left(G((1-\beta^n)m_t^{(i)} - \delta_-, (1-\beta^n)m_t^{(i)} + \delta_+, n\sigma^2 + (1-\beta^{2n})\omega^2) \mid m_t^{(i)} > m_t^{(\neq i)}\right), \\
&> \frac{n\sigma^2 + \beta^{h-n}(1-\beta^{2n})\omega^2}{n\sigma^2 + (1-\beta^{2n})\omega^2} G((1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N - \delta_-, (1-\beta^n)\omega\sqrt{1-\rho}\,\psi_N + \delta_+, n\sigma^2 + (1-\beta^{2n})\omega^2).
\end{aligned} \tag{29}$$

In the first step, I use that the conditional expectation of ε_{t+j} and η_{t+j} is zero for $j > n$. In the second step I use the result in Eq. (26), and the final step follows from Jensen's inequality and the convexity of G. Taking the limit $h \to \infty$ eliminates the β^{h-n} term because $0 \leq \beta < 1$. Collecting terms yields the expression in Eq. (20). For one-sided last look with $\delta_+ = \infty$, the expression in Eq. (28) constitutes an upper-bound on the conditional expectation of $m_t^{(i)}$. Consequently the expression in Eq. (20) provides a lower-bound for the effective spread for one-sided last look.

Next, applying the on-latest last look protocol as in Definition 4 to the definition of the effective spread in Eq. (19) gives:

$$
\begin{aligned}
\mathbb{S} &= s + 2E\left(p_{t+n}^{*} - p_{t+n}^{(i)} \mid b_t^{(i)} > b_t^{(\neq i)}, -\delta_- < b_{t+n}^{(i)} - b_t^{(i)} < \delta_+\right), \\
&= s - 2E\left(m_{t+n}^{(i)} \mid b_t^{(i)} > b_t^{(\neq i)}, -\delta_- < m_{t+n}^{(i)} - m_t^{(i)} + p_{t+n}^{*} - p_t^{*} < \delta_+\right), \\
&= s - 2E\left(\beta^n m_t^{(i)} + \sum_{j=0}^{n-1} \beta^j \eta_{t+n-j}^{(i)} \mid b_t^{(i)} > b_t^{(\neq i)}, -\delta_- < m_{t+n}^{(i)} - m_t^{(i)} + p_{t+n}^{*} - p_t^{*} < \delta_+\right), \\
&\approx s - 2\beta^n \omega \sqrt{1-\rho}\,\psi_N - 2(1-\beta^{2n})\frac{\omega^2}{\sigma_\alpha^2} G(\mu_\alpha - \delta_-, \mu_\alpha + \delta_+, \sigma_\alpha^2).
\end{aligned}
$$

This completes the proof of Theorem 1. ■

References

Amano, K., N. Goda, S. Nishida, Y. Ejima, T. Takeda, and Y. Ohtani, 2006, "Estimation of the timing of human visual perception from magnetoencephalography," *Journal of Neuroscience*, 26 (15), 3981 – 3991.

Bank of England, H.M. Treasury, and Financial Conduct Authority, 2015, "How fair and effective are the fixed income, foreign exchange and commodities markets?," available at http://www.bankofengland.co.uk/markets/Documents/femrjun15.pdf.

Budish, E., P. Cramton, and J. Shim, 2015, "The high-frequency trading arms race: frequent batch auctions as a market design response," *Quarterly Journal of Economics*, 130 (4), 1547 – 1621.

Cartea, A., S. Jaimungal, and J. Walton, 2015, "Foreign exchange markets with last look," working paper, University College London.

Cliff, D., D. Brown, and P. Treleaven, 2010, "Technology trends in the financial markets: a 2020 vision," UK Government's Foresight Project, available at https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/289029/11-1222-dr3-technology-trends-in-financial-markets.pdf.

Foucault, T., A. Röell, and P. Sandås, 2003, "Market making with costly monitoring: An analysis of the SOES Controversy," *Review of Financial Studies*, 16 (2), 345 – 384.

Harris, J. H., and P. H. Schultz, 1998, "The trading profits of SOES bandits," *Journal of Financial Economics*, 50, 39 – 62.

Kyle, A. S., 1985, "Continuous auctions and insider trading," *Econometrica*, 53 (6), 1315–1335.

Norges Bank Investment Management, 2015, "The role of last look in foreign exchange markets," The Asset Manager Perspective series, available at http://www.nbim.no/en/transparency/asset-manager-perspectives.

Oomen, R. C., 2016, "Execution in an aggregator," forthcoming *Quantitative Finance*.

Stuart, A., and K. Ord, 1994, *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Wiley-Blackwell, London, UK, 6 edn.

UNCITRAL, 2010, "United Nations Convention on Contracts for the International Sale of Goods," United Nations Commission on International Trade Law Text, available at http://www.uncitral.org/uncitral/en/uncitral_texts/sale_goods.html.