Finance and Stochastics (2022) 27:1–47
https://doi.org/10.1007/s00780-022-00491-w

Optimal execution with stochastic delay

Álvaro Cartea[1,2] **· Leandro Sánchez-Betancourt**[3]

Received: 1 June 2022 / Accepted: 6 July 2022 / Published online: 1 December 2022

Abstract
We show how traders use marketable limit orders (MLOs) to liquidate a position over a trading window when there is latency in the marketplace. MLOs are liquidity-taking orders that specify a price limit and are for immediate execution only; however, if the price limit of the MLO precludes it from being filled, the exchange cancels the order. We frame our model as an impulse control problem with stochastic latency where the trader controls the times and the price limits of the MLOs sent to the exchange. We show that impatient liquidity takers submit MLOs that may walk the book (capped by the limit price) to increase the probability of filling the trades. On the other hand, patient liquidity takers use speculative MLOs that are only filled if there has been an advantageous move in prices over the latency period. Patient traders who are fast do not use their speed to hit the quotes they observe, or to finish the execution programme early; they use speed to complete the execution programme with as many speculative MLOs as possible. We use foreign exchange data to implement the random-latency-optimal strategy and to compare it with four benchmarks. For patient traders, the random-latency-optimal strategy outperforms the benchmarks by an amount that is greater than the transaction costs paid by liquidity takers in foreign exchange markets. Around news announcements, the value of the outperformance is between two and ten times the value of the transaction costs. The superiority of the strategy is due to both the speculative MLOs that are filled and the price protection of the MLOs.

Keywords Algorithmic trading · High-frequency trading · Stochastic delay · Latency

Mathematics Subject Classification (2020) 93E20 · 91G80 · 49L20 · 49L25

JEL Classification C02 · C61

✉ Á. Cartea
alvaro.cartea@maths.ox.ac.uk

L. Sánchez-Betancourt
leandro.sanchez-betancourt@kcl.ac.uk

1 Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK

2 Oxford-Man Institute of Quantitative Finance, Oxford, OX2 6ED, UK

3 Department of Mathematics, King's College London, London, WC2R 2LS, UK

1 Introduction

In electronic trading, agents are exposed to the risks that stem from latency in the marketplace. Latency is the aggregate of the time lags associated with the various stages of a trade. These stages, which occur in succession and are separated in time by random delays, include: the exchange sends quotes to an agent, the agent receives the quotes, the agent processes information and sends an instruction to the exchange, the exchange receives and handles the instruction, and finally, the outcome is notified to the agent. During the latency period, it is likely that the exchange will process other instructions that modify the limit order book (LOB) and possibly affect the outcome of the agent's instructions. In particular, liquidity takers face the risk that the prices they target are not available by the time the exchange processes their order because the best quotes they observed were stale and therefore updated during the latency period. The risks faced by liquidity providers are different; e.g. their limit orders may be adversely selected by a taker. If the market moves against the trader's interest, the order will not be filled or will be filled at worse prices – the outcome depends on the type of liquidity-taking order sent by the trader.

At high frequencies (milliseconds, microseconds), the best quotes in the LOB tend to flicker. Flickers are unpredictable short-lived deviations of a few ticks from the best quotes, and are the result of very frequent occurrences of rapid sequences of post-and-cancel activity in the LOB and the less frequent arrival of aggressive orders that consume liquidity that is immediately replenished. Thus over the latency period, the main source of risk faced by liquidity takers is from flickers in the book, and, to a lesser extent, the risk that stems from unexpected changes in the fundamental best quotes by the time their orders are processed. If latency were zero, traders would not face these risks because they would always take liquidity at the prices and quantities they observe in the LOB; however, all market participants face latency and latency is random.

In this paper, we show how a trader can execute a large position in a financial instrument when the trader faces random latency in the marketplace. The dynamics of the best bid price in the LOB consist of the 'fundamental' best quote of the instrument and the flickers (similarly for the best ask price). Innovations in the fundamental best quote are driven by a stochastic process, and the size and arrival times of flickers in the quotes are represented by a marked point process. The trader employs marketable limit orders (MLOs) to execute the position over a time window; MLOs are liquidity-taking orders that specify a price limit and are for immediate execution only. Filled MLOs have permanent price impact: a filled buy (sell) MLO exerts upward (downward) pressure on the fundamental value of the instrument. However, if the price limit of the MLO precludes it from being filled, the exchange cancels the order and there is no price impact because other market participants cannot observe missed trades.

In our model, the trader controls the price limit of the MLO and controls when to submit it to the exchange, both of which largely depend on the trader's degree of urgency to execute the position. An impatient liquidity taker will send sell (buy) MLOs with price limits that are below (above) the best bid (ask) price they observed in the LOB. This strategy increases the probability that the MLO is filled when processed

by the exchange and caps how far the MLO is allowed to walk the LOB when there are adverse price changes over the latency period. On the other hand, we show that patient liquidity takers use MLOs predominantly to send speculative trades. Speculative MLOs seek a price improvement relative to the fundamental best quote observed by the trader when sending the order to the exchange. Patient liquidity takers do not use their speed to finalise their execution programme ahead of time, or to hit the best quotes they observe in the LOB, they use speed mainly to send as many speculative MLOs as possible during the trading horizon – fast traders have many opportunities to retry missed orders before reaching the end of the trading window.

We use proprietary foreign exchange data from the LMAX Exchange (henceforth LMAX) between September 2019 and February 2020 for ten currency pairs to study the order types that liquidity takers use in the foreign exchange market and to analyse the effect of latency on the efficacy (i.e., hit and miss) of liquidity-taking orders. In all pairs, MLOs trade more volume than any other type of liquidity-taking order; in particular, more than market orders (MOs), which are for immediate execution and walk the LOB until filled in full. Though MLOs offer protection against adverse price moves, traders concede that the order may be cancelled by the exchange because clearing prices would breach the price limit instructed in the MLO. For example, in the EUR/USD pair, 41.0% of the MLOs were filled in full, 1.9% were partially filled, and 57.1% were missed. The MLOs that were partially or fully filled represent 56.9% of the volume traded in the pair, while MOs represent 13.5% of the total volume. The limit rates of missed MLOs were, on average, three ticks away from the best quote in the LOB when they were processed by the exchange.

We use the LMAX data to implement and to benchmark the performance of the trader's random-latency-optimal execution strategy against four strategies: (i) send MOs over a trading window and assume latency is zero, (ii) send MLOs over a trading window and assume that latency is deterministic, (iii) employ a discrete time-weighted average price (TWAP) strategy that sends MOs at equally spaced time intervals over the execution window, (iv) send one MO to execute the entire order at the beginning of the trading window, where, unrealistically, we assume that there is enough liquidity at the best bid in the LOB to fill the entire order.

Estimates of model parameters for the EUR/USD currency pair are obtained with data between 1 August 2019 and 31 August 2019, and trading strategies are implemented with data between 1 September 2019 and 29 February 2020. We show that the performances of the random-latency-optimal strategy and of the deterministic-latency-optimal strategy are statistically the same. When the trader is patient, both strategies outperform the other benchmarks by a cash amount that is greater than the transaction costs paid by liquidity takers in foreign exchange markets, and around news events, the value of the outperformance increases to between two and ten times the value of the transaction costs. The source of the better performance of the latency-optimal strategies stems from the speculative MLOs that are filled during the trading interval, and from the price protection of the MLOs against adverse price moves. In the EUR/USD pair, the number and the value of filled speculative MLOs increase when market activity increases because the probability and the size of positive flickers in the best bid rate increase during heightened market activity – this explains the considerable outperformance of the latency-optimal strategies around news events.

Finally, we show that the value of the outperformance decreases as the degree of impatience of the trader increases because the strategy sends fewer speculative MLOs – very impatient traders do not send speculative MLOs, all their MLOs are for price protection.

As far as we are aware, this is the first work to appear in the literature that shows how to optimally execute a position in an asset when the trader faces random latency. Closest to our work are those by Øksendal and Sulem [25] and Bruder and Pham [9], where the authors investigate general impulse control problems with deterministic delay, while ours assumes stochastic delay; [25] studies an infinite-horizon control problem with deterministic delay and an arbitrary number of pending orders, while [9] investigates a finite-horizon stochastic control problem with any finite number of pending orders. The key mathematical novelty that distinguishes our work is that delays are stochastic. From a modelling perspective, this requires an explicit definition of the σ-algebra of the trader and an account of the new terms that arise as a consequence of stochastic delays.

Although our model focuses on an optimal execution problem with stochastic latency, our framework is general and can be applied to control problems in which the outcomes of the controlled actions are observed with stochastic delay. In financial applications, our model may be implemented in all asset classes that trade in electronic LOBs with liquidity-taking orders that consider price limits and immediate time-in-force of the liquidity-taking orders (equity and foreign exchanges that offer MLOs include LMAX, LSE, Chi-X, CBOE, NASDAQ, CME, and NYSE). In contrast, all models of optimal execution in the extant literature assume that the trader operates in the marketplace with zero latency and employ only MOs to take liquidity; see e.g. Almgren and Chriss [4], Almgren [2], Bayraktar and Ludkovski [6], Alfonsi et al. [1], Guéant et al. [21], Guilbaud and Pham [22], Cartea et al. [13], Cartea and Jaimungal [10], Guéant [19], Cartea and Jaimungal [11], Barger and Lorig [5] and the monographs by Cartea et al. [12, Sect. 6] and Guéant [20, Part II].

In the literature, there is work that discusses how latency affects liquidity-taking strategies, market making and passive trading. Cartea and Sánchez-Betancourt [15] show how traders can adjust the price limit of MLOs to target a fill ratio in a given currency pair when the trader sends a sequence of orders over a trading interval (finite or infinite). In market making, Gao and Wang [18] show how an agent provides liquidity to the LOB of large-tick stocks when latency is deterministic and fixed during the trading interval. In passive trading (i.e., trading with limit orders), the work by Moallemi and Saĝlam [24] quantifies the cost of latency in equity (NASDAQ).

In the next section, we employ high-frequency foreign exchange data to show that MLOs protect orders from adverse price moves and also receive price improvements when prices move in the trader's interest over the latency period. Section 3 introduces the trader's optimal execution problem with random latency and characterises the random-latency-optimal strategy as the solution of a Hamilton–Jacobi–Bellman quasi-variational inequality (HJBQVI), and develops two new benchmarks where the investor faces deterministic latency or faces no latency. Section 4 compares the performance of the trader's random-latency-optimal strategy with that of the benchmarks. Section 5 concludes, and we collect some proofs in the appendices.

2 Order types and data

2.1 Orders and revealed preferences

In order-driven electronic markets, the basic building blocks to trade are orders that provide liquidity and orders that take liquidity. When a trader sends an order to the exchange, the order contains two specific instructions: quote type and time-in-force. The quote type specifies the main feature of the order: price limit (i.e., LO), no price limit (i.e., MO), a trigger (i.e., stop orders), among others. The time-in-force refers to how long the order is active in the market. The shortest time-in-force is immediate execution and the longest time-in-force is typically for the rest of the trading day. Two prominent liquidity-taking order types that have an immediate execution time-in-force are immediate-or-cancel (IoC) and fill-or-kill (FoK). IoC is an order to buy or to sell assets that must be executed immediately, in full or in part, while obeying the order's price limit, and any portion of the volume of the order that cannot be filled at the desired price limit is cancelled. FoK is an order to buy or to sell assets that must be executed immediately in full or it is cancelled.

Combinations of the various features constitute the type of order that market participants send to the exchange. Here, we focus on the types of liquidity-taking order that are designed to protect traders against the frictions that stem from latency. We define MLOs as liquidity-taking orders that have a price limit and the time-in-force is IoC or FoK. An MO is an MLO without price limit; the MLOs we consider have a finite price limit and are for immediate execution. MLOs protect liquidity takers from adverse price movements that occur between the time the trader makes a decision to trade (on a possibly stale quote) and the time the exchange matches the order with a limit order resting in the book (when possible).

Traders reveal their preferences when they choose a type of liquidity-taking order for a trade or a sequence of trades whose outcome is contingent on latency. Their choices demonstrate that traders balance the cost of completing a trade and the costs of price protection. A trader who must complete a trade without delay will choose an MO to guarantee execution in full – and expose the trade to price movements over the latency period. In contrast, a trader who can afford to miss the trade if the price is 'not right' will choose an MLO. Thus in exchange for price protection against adverse price movements, the trader concedes that the trade might not get filled. On the other hand, if prices move in a favourable direction, the trader receives a price improvement, i.e., the MLO is executed at a better price than the trader's decision price. Indeed, MLOs are also used by traders to complete a trade only if the price improves over the latency period. In the next section, we provide summary statistics of the use of MLOs in foreign exchange markets.

2.2 Data

We present descriptive statistics for ten currency pairs in the foreign exchange spot market at LMAX. The data are stamped at a microsecond frequency and the range is from 1 September 2019 to 29 February 2020 – in Sect. 4, we use data from August 2019 to estimate model parameters. For each currency pair, we have: the aggregate

Table 1 Percentage of liquidity-taking by order type. Period: 1 September 2019 to 29 February 2020. For each currency pair, we use bold to highlight the highest percentage by number of trades and by volume traded

	Number of trades			Volume traded		
	MLOs	MOs	Other	MLOs	MOs	Other
EUR/USD	**36.7%**	32.1%	31.2%	**56.9%**	13.5%	29.6%
USD/JPY	41.4%	6.9%	**51.7%**	**55.8%**	3.1%	41.1%
EUR/GBP	38.3%	8.0%	**53.7%**	**73.0%**	5.2%	21.8%
GBP/USD	**53.6%**	21.9%	24.5%	**72.8%**	7.1%	20.2%
USD/MXN	**44.7%**	24.6%	30.7%	**52.0%**	26.3%	21.7%
USD/CAD	**62.2%**	10.9%	26.9%	**71.2%**	2.6%	26.1%
GBP/JPY	**49.1%**	19.4%	31.5%	**61.2%**	8.5%	30.3%
EUR/JPY	33.0%	7.3%	**59.8%**	**65.0%**	4.3%	30.8%
AUD/USD	**66.7%**	10.6%	22.7%	**71.0%**	4.4%	24.6%
AUD/JPY	**52.8%**	11.7%	35.5%	**59.9%**	6.1%	34.0%

volume of the limit sell orders posted at the best ask rate; the aggregate volume of the limit buy orders posted at the best bid rate; liquidity-taking orders sent to the LOB, including the type of order and trader identification; rate limits of MLOs; full fills, partial fills, and missed liquidity-taking trades. Finally, our data set does not contain the liquidity posted at the LOB beyond the best bid and best ask rates, but does contain the average rate paid or received for all liquidity-taking orders, including those that walked the LOB beyond the best quotes.

Table 1 shows the percentage of the total number of trades that are MLOs, MOs and others, and also shows the percentages by traded volume. Recall that the MLOs we consider are IoC and FoK with an immediate time-in-force. In practice, the exchange processes the order as soon as it arrives and matches it (if possible). This processing time usually takes under 80 microseconds and is followed by a message from the exchange to the trader to notify the outcome.

The category 'others' includes the remaining order types that consume liquidity, for example, marketable orders: (i) limit orders that are good-until-cancel and good-for-day, (ii) stop orders good-for-day and good-until-cancel, and (iii) dark limit orders good-for-day and good-until-cancel.

2.3 Flickering of best rates in the LOB

Figure 1 shows snippets of the evolution of best quotes and liquidity-taking activity for the currency pair EUR/USD on 3 September 2019 at around 2.00pm British summer time – the data are timestamped to the microsecond. Solid circles denote filled sell MLOs, squares denote filled sell MOs, and empty circles denote sell MLOs that were not filled because the limit sell rate in the order was higher than the best bid rate in the LOB when the exchange processed the order.

The top and middle panels of Fig. 1 show the best bid rate and sell liquidity-taking orders between 2.00.00pm and 2.01.00pm (i.e., one minute) and between 2.00.08pm and 2.00.14pm (i.e., six seconds), respectively. It is clearly visible that the best bid rate goes through unpredictable flickers, which are extremely short-lived deviations of a significant number of ticks from the 'fundamental' or 'true' value of the best bid

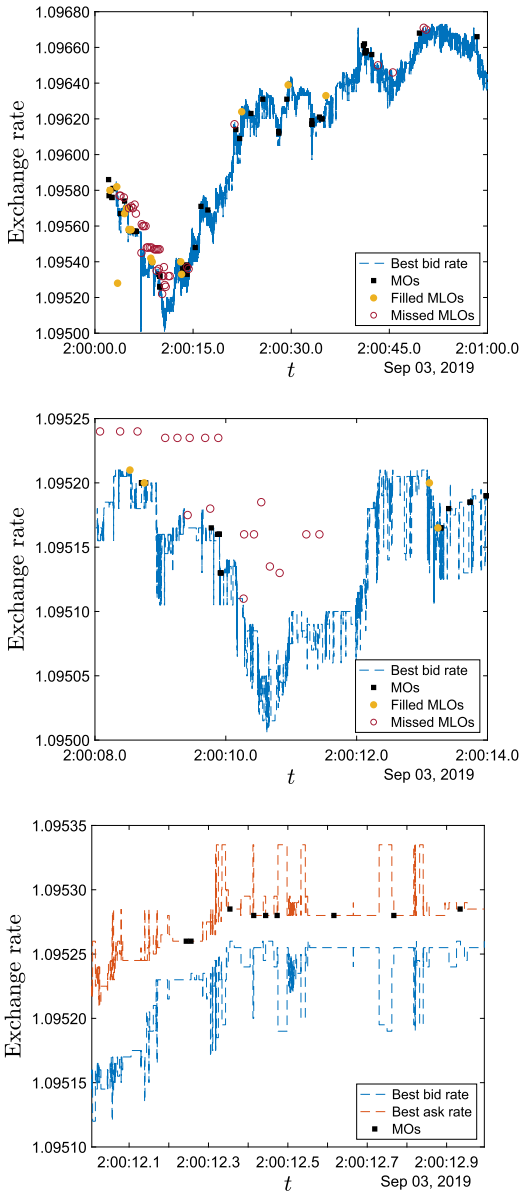

Fig. 1 Best bid rate (blue line), best ask rate (red line), missed MLOs (empty circles), filled MLOs (full circles), filled MOs (squares). Top panel: 2.00pm to 2.01pm (60 seconds). Middle panel: 2.00.08pm to 2.00.14pm (6 seconds). Bottom panel: 2.00.12pm to 2.00.13pm (1 second). Microsecond data from EUR/USD pair, 3 September 2019

rate. These flickers are the result of cancellations and arrivals of limit orders at the best bid rates, and of liquidity that is consumed by aggressive orders and immediately replenished by the arrival of limit orders.

The bottom panel shows a one-second snippet, between 2.00.12pm and 2.00.13pm, of the best bid rate and the best ask rate and the liquidity-taking orders, all of which were buy MOs. In most cases, flickers on either side of the LOB cause a short-lived widening of the quoted spread of the LOB.

From visual inspection, flickers in the best bid (best ask) rate are more often negative (positive) than positive (negative). This is a prevalent feature in the dynamics of the best quotes in the LOB. We return to this in Sect. 4 when we estimate model parameters and discuss the distribution of the flickers in the performance of the execution strategies.

2.4 Distribution of flickers: hit and miss

We proceed to discuss the limit rates employed by traders who send MLOs to buy and to sell the EUR/USD currency pair. We employ the same data as that in Table 1 – the results for the remaining currency pairs are similar. To analyse both the hit and miss performance of the MLOs and the distribution of the flickers in the best quotes, we compute the slippage-price-improvement (SPI) measure

$$\text{SPI}_i = (L_i - M_i)I_i. \tag{2.1}$$

Here, L_i denotes the exchange rate limit of the order, i.e., the maximum (minimum) rate that the trader is willing to pay (receive) per unit that she wishes to buy (sell); M_i is the exchange rate that the order would pay (receive) per unit bought (sold) if the order is filled in full. The indicator I_i determines the direction of the order: I_i takes the value $+1$ when the trader sends a buy MLO and the value -1 when the trader sends a sell MLO, for $i = 1, \ldots, n$, where n is the total number of MLOs. Thus when SPI_i is positive, the order was filled (relative to the limit rate L_i) with a slack of SPI_i ticks, which we refer to as price improvement. Similarly, when SPI_i is negative, the order missed (fully or in part) by SPI_i ticks (relative to the limit rate L_i), which we refer to as slippage.

For each MLO, the quantity in (2.1) could understate or overstate slippage and price improvement relative to the trader's decision rate; note that we compute SPI relative to the limit rate instructed in the order. Our data set does not contain the time-stamps of when the trader submitted the MLO to the exchange, nor does it contain the rate observed by the trader when she decided to trade (i.e., the decision rate). We have the time-stamp of when the order is processed by the exchange. Thus we do not know the best bid and best ask rates and quantities posted in the LOB when the trader decided to send the order.

The trader may find it optimal to send a buy MLO with a limit rate that is below the observed best ask rate or to send a sell MLO with a limit rate that is above the observed best bid rate. In other words, traders may send speculative MLOs that are filled only if there is a price improvement relative to the quotes they observe when deciding to trade – in Sect. 3.6 below we return to this point and show that it is optimal to send MLOs that target a price improvement.

Figure 2 shows histograms in log-scale of SPI for the MLOs in the EUR/USD currency pair, where the x-axes in both panels are truncated to lie between -25 and 25 ticks. To gain insights into the limit rates of the MLOs and into the distribution of the flickers, the top panel shows the histogram of SPI with all the MLOs sent to the exchange, and the bottom panel shows a histogram of SPI with the MLOs that aimed at the best quote, i.e., the trader's decision rate and the limit rate she instructed in the MLO are the same. We consider that the limit rate of an MLO is equal to the decision rate of the trader if at any point within the previous 150 ms of the processing time of

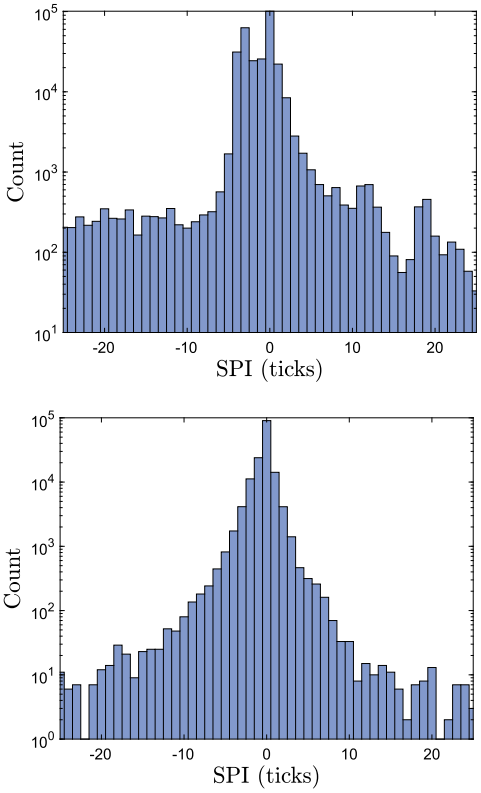

Fig. 2 Histograms of SPI between −25 ticks and 25 ticks in log-scale for the EUR/USD, 1 September 2019 to 29 February 2020. One tick is 10^{-5} USD. Top: all MLOs. Bottom: MLOs with limit rate equal to trader's decision rate

the MLO, there was a best quote in the LOB equal to the limit rate of the MLO (we obtain similar results when we assume that the time period of the look-back window is 100 ms). The histograms are for both IoCs and FoKs.

In both panels, the histograms are skewed to the left. The skewness may be due to informed traders in the market and due to the distribution of the flickers. Informed traders send buy (sell) MLOs in anticipation of an increase (decrease) in the exchange rate. Thus all else being equal, their trading strategies skew the distribution of SPI to the left because it is more likely that a buy (sell) MLO misses the trade when the exchange rate drifts up (down). Informed MLOs are included in the top panel. In addition, the skew in the histogram also results from asymmetry in the distribution of the size of the flickers – this is better represented in the bottom panel where the limit rate of the MLO is equal to the trader's decision rate; see also the bottom panel of Fig. 1, where most flickers on the best bid (ask) rate are negative (positive). Below we show that flickers in the best bid rate are negatively skewed and flickers in the best ask rate are positively skewed.

2.5 Liquidity: make and take

In this subsection, we report various statistics of the volume posted at the best quotes in the LOB and the volume of the filled liquidity-taking orders of the ten currency

Table 2 TWS: time-weighted spread. TWQ: time-weighted quantity at best quotes. Period: 1 September 2019 to 29 February 2020 between 9.00am and 4.00pm. Each unit of quantity is a lot of 10'000 units of the base currency

	Liquidity-taking (volume)			Liquidity provision			
Pair	Mean	Std	Median	TWS ask/bid	TWQ ask/bid	Median Q ask/bid	Mode Q
EUR/USD	20.94	26.64	4.00	3.94×10^{-5}	160/159	100/100	50/50
USD/JPY	26.81	23.78	27.00	4.20×10^{-3}	110/109	100/100	50/50
EUR/GBP	12.93	21.50	0.60	9.30×10^{-5}	94/94	80/82	50/50
GBP/USD	20.65	21.97	14.40	9.83×10^{-5}	91/91	75/66	50/50
USD/MXN	37.09	23.55	50.00	4.67×10^{-3}	80/80	50/50	50/50
USD/CAD	31.30	27.42	30.00	8.34×10^{-5}	101/104	100/100	50/50
GBP/JPY	20.03	16.35	30.00	1.55×10^{-2}	71/71	50/50	50/50
EUR/JPY	14.44	22.33	0.60	8.77×10^{-3}	100/100	100/100	50/50
AUD/USD	32.77	28.23	32.10	5.65×10^{-5}	113/114	100/100	50/50
AUD/JPY	31.08	23.27	40.00	8.62×10^{-3}	88/90	94/99	50/50

Table 3 TWQ: time-weighted quantity available at best quotes. TQ: traded quantity. Q ask/bid: quantity posted at best ask rate and best bid rate. Period: 1 September 2019 to 29 February 2020 between 9.00am and 4.00pm. Each unit of quantity is for 10'000 units of the base currency

Currency pair	TWQ ask/bid	TQ > TWQ	Median Q ask/bid	TQ > MQ	Mode Q ask/bid	TQ > MoQ
EUR/USD	160/159	0.1%	100/100	0.2%	50/50	4.6%
USD/JPY	110/109	0.1%	100/100	0.1%	50/50	2.6%
EUR/GBP	94/94	1.2%	80/82	1.4%	50/50	1.9%
GBP/USD	91/91	1.2%	75/66	1.5%	50/50	2.3%
USD/MXN	80/80	4.6%	50/50	5.9%	50/50	5.9%
USD/CAD	101/104	0.2%	100/100	0.2%	50/50	7.6%
GBP/JPY	71/71	0.7%	50/50	0.9%	50/50	0.9%
EUR/JPY	100/100	1.8%	100/100	0.2%	50/50	2.8%
AUD/USD	113/114	0.1%	100/100	0.1%	50/50	8.8%
AUD/JPY	88/90	1.5%	94/99	1.4%	50/50	0.1%

pairs we study. The key message is that liquidity-taking orders hardly ever consume all the liquidity available at the best quotes in the LOB.

Table 2 reports statistics of the liquidity-taking and the liquidity-provision activity during the period 1 September 2019 to 29 February 2020 between 9.00am and 4.00pm British summer time. Columns 2–4 describe the liquidity-taking activity. The mean, median and standard deviation are in 10'000 units of the base currency. For example, the mean value of the traded quantity in the pair EUR/USD is 20.94, which, in the base currency EUR, is a mean of 209'400 EUR. The last four columns of the table summarise statistics of the spread and of the best quotes in the LOB: the time-weighted spread for each currency pair, the time-weighted quantity, the median quantity available at the best quotes, and the mode quantity at the best quotes.

Table 3 shows the percentage of occurrences when the traded quantity is greater than: the time-weighted quantity at the best quotes; the median quantity at the best quotes; and the mode of the quantity at the best quotes. Note that the information in columns 1, 3, 5 is the same as that in the last three columns of Table 2.

We observe that in all currency pairs, the proportion of trades that cannot be filled with the liquidity available at the best quotes is between 0.1% and 8.8%. In the model that follows, we assume that the MLOs sent by the trader to the exchange will not walk the LOB.

3 Optimal execution with random latency

We present a model of optimal execution with random latency. We focus on the execution of a large order that is divided in child orders that are sent to the market over a finite trading horizon. However, we remark that our model is useful to decide the optimal strategy to execute as little as one order over a trading horizon. Our discussion is framed within foreign exchange markets; however, our model is general and applicable in all asset classes in which instruments are traded in a visible electronic LOB. Also, our framework may be applied in general stochastic control problems in which the outcomes of the agent's actions are known at a random future date.

3.1 Exchange rate dynamics and execution of MLOs

As discussed above, flickers are short-lived deviations; so here we assume that only the flicker that occurs at the processing time $\tilde{t}_0$ is relevant when the exchange processes the trader's sell MLO. Therefore, the trader models the best bid rate $\hat{S} = (\hat{S}_t)_{t\geq 0}$ as

$$\hat{S}_t = S_t + F_t, \tag{3.1}$$

where $S = (S_t)_{t\geq 0}$ denotes the fundamental best bid rate, and $F = (F_t)_{t\geq 0}$ denotes the flickers that affect the bid rate only at processing times of MLOs. The model (3.1) does not include flickers that occur between processing times of the MLOs sent by the trader – it models paths that are relevant for the trader. In the sequel, we refer to $\hat{S}_t$ as the observed best bid rate, and in Sect. 4, we show how the trader employs the data of the LOB to obtain the fundamental best bid rate S_t.

Formally, the notification time of the order and the size of the flicker are modelled by the background marked point process (MPP) $\mathcal{N} = (T_n, Z_n)_{n\geq 1}$ with random measure $\mathfrak{p}(\mathrm{d}t, \mathrm{d}z)$. The background MPP dictates the stopping time when the outcome of each order is notified to the trader and dictates the flicker that affects the fundamental bid rate when each order is processed. That is, the trader sends a sell MLO to the exchange at time t and the exchange notifies the outcome of the order at time $\tilde{t}$, where $\tilde{t}$ is the next stopping time (after t) when a mark (i.e., flicker) of the MPP arrives. More precisely, let $n \geq 1$ be such that $T_{n-1} \leq t < T_n$; then the notification time $\tilde{t}$ is T_n and the flicker $F_{\tilde{t}}$ is Z_n. Note that the value $F_{\tilde{t}}$ of the flicker at the notification time is independent of the trader's information at time t.

We assume that the background process $\mathcal{N}$ is non-explosive and the interarrival times $(T_i - T_{i-1})_{i\geq 1}$ are independent and identically distributed exponential random variables with parameter $\lambda > 0$ and $T_0 = 0$. The flickers $(Z_n)_{n\geq 1}$ that affect the fundamental best bid rate at processing times take values in $\mathbb{R}$ and are i.i.d. with law

$$\nu(\mathrm{d}z) = p_0\delta_{\{0\}}(\mathrm{d}z) + \underbrace{p_+\eta_+ e^{-\eta_+ z}\mathbf{1}_{\{z>0\}}\,\mathrm{d}z}_{\text{price improvement}} + \underbrace{p_-\eta_- e^{\eta_- z}\mathbf{1}_{\{z<0\}}\,\mathrm{d}z}_{\text{slippage}}, \tag{3.2}$$

where $\delta_{\{0\}}(\mathrm{d}z)$ is the Dirac measure at zero, $\eta_+, \eta_- \in (0,\infty)$ and $p_0, p_+, p_- \in \mathbb{R}_+$ with $p_0 + p_+ + p_- = 1$.

Let $T \in (0,\infty)$. It is straightforward to see that

$$\mathbb{E}\big[\big(\mathfrak{p}([0,T],\mathbb{R})\big)^2\big] < \infty \qquad \text{and} \qquad \mathbb{E}\left[\int_0^T\int_{\mathbb{R}} |z|\,\mathfrak{p}(\mathrm{d}t,\mathrm{d}z)\right] < \infty. \tag{3.3}$$

The law in (3.2) is a modelling choice; the framework we develop here is valid for any law $\nu(\mathrm{d}z)$ that satisfies the two conditions in (3.3).

3.2 Outcome of trade attempts: miss or fill

The information flow that the trader observes is encoded in the filtration $\mathbb{F} = (\mathcal{F}_t)_{t\geq 0}$ defined by

$$\mathcal{F}_t = \sigma\big(W_s, \mathfrak{p}([0,s],A) : A \in \mathcal{B}(\mathbb{R}), s \leq t\big),$$

where $(W_s)_{s\geq 0}$ is a standard Brownian motion. Let τ be an $\mathbb{F}$-stopping time, and set $N_t = \mathfrak{p}((0,t],\mathbb{R})$ for $t \in (0,T]$ and $N_0 = 0$, where N_t denotes the total number of marks of the background process $\mathcal{N}$ up to time t. We define the notification time $\tilde{\tau}$ of a trade attempt by

$$\tilde{\tau} = \inf\{t > \tau : \Delta N_t > 0\},$$

where $\Delta N_t = N_t - N_{t-}$. Thus $\tilde{\tau}$ denotes the time when the exchange notifies the outcome of the trade attempt sent at time $\tau < \tilde{\tau}$. **Henceforth**, we put a tilde $\tilde{}$ on stopping times to denote that they are notification times.

The following lemma shows that: (i) the notification times are stopping times, and (ii) the times between trade attempts and notification times are i.i.d. with exponential distribution.

Lemma 3.1 *Let $(\tau_i)_{i\in\mathbb{N}}$ be a sequence of $\mathbb{F}$-stopping times, where the index i denotes the ith trade attempt, and let $(\tilde{\tau}_i)_{i\in\mathbb{N}}$ be the sequence of notification times of the outcome of each trade attempt. Let $(\tau_i)_{i\in\mathbb{N}}$ and $(\tilde{\tau}_i)_{i\in\mathbb{N}}$ satisfy $\tilde{\tau}_i \leq \tau_{i+1}$ for all $i \in \mathbb{N}$. Then*:

(i) *$(\tilde{\tau}_i)_{i\in\mathbb{N}}$ are $\mathbb{F}$-stopping times.*

(ii) *$(\tilde{\tau}_i - \tau_i)_{i\in\mathbb{N}}$ constitute a collection of independent and identically distributed random variables that are exponentially distributed with parameter $\lambda > 0$.*

Proof (i) Let $j \in \mathbb{N}$ and $t \geq 0$. Then

$$\{\tilde{\tau}_j \leq t\} = \{\tau_j \leq t\} \cap \{N_t > N_{t \wedge \tau_j}\} \in \mathcal{F}_t$$

because $\{\tau_j \leq t\} \in \mathcal{F}_t$ and $\{N_t > N_{t \wedge \tau_j}\} \in \mathcal{F}_t$.

(ii) For $i \in \mathbb{N}$ and $t \geq 0$, we have

$$\begin{aligned}\mathbb{P}[\tilde{\tau}_i - \tau_i \geq t] &= \mathbb{P}[\inf\{s > 0 : \Delta N_{s+\tau_i} > 0\} \geq t] \\ &= \mathbb{P}[\inf\{s > 0 : \Delta N_s > 0\} \geq t] = e^{-\lambda t}\end{aligned}$$

because $(N_{\tau_i+s} - N_{\tau_i})_{s \geq 0}$ is a Poisson process with parameter λ. Finally, the collection $(\tilde{\tau}_i - \tau_i)_{i \in \mathbb{N}}$ has the independence property because $[\tau_i, \tilde{\tau}_i)_{i \in \mathbb{N}}$ are over non-overlapping intervals. □

Next, we define the auxiliary process $\mathsf{F} = (\mathsf{F}_t)_{t \geq 0}$ to determine the outcome of the trade attempts. The value of F_t is the most recent mark up to time t. Thus F satisfies

$$\mathrm{d}\mathsf{F}_t = \int_{\mathbb{R}} (z - \mathsf{F}_{t-}) \mathfrak{p}(\mathrm{d}t, \mathrm{d}z), \qquad \mathsf{F}_0 = 0,$$

where we recall that $\mathfrak{p}(\mathrm{d}t, \mathrm{d}z)$ is the random measure of the background process $\mathcal{N}$. When the trader sends a trade at time τ and the notification time is $\tilde{\tau}$, the value $F_{\tilde{\tau}}$ of the flicker is $\mathsf{F}_{\tilde{\tau}}$.

At the notification time $\tilde{\tau}$, if the sell MLO is an FoK with limit rate $\mathfrak{l}$, it will be filled in full if $\mathfrak{l} \leq S_{\tilde{\tau}-} + \mathsf{F}_{\tilde{\tau}}$; otherwise the MLO misses the trade and is cancelled by the exchange. To formally define a miss and a fill of a trade attempt, we proceed as follows. Let $\tilde{H}(x) = \mathbf{1}_{\{x \leq 0\}}$ be a step function. The outcome of the sell MLO is determined by the step function $\tilde{H}(\mathfrak{l} - S_{\tilde{\tau}-} - \mathsf{F}_{\tilde{\tau}})$, which takes the value 1 when the trader receives the notification of a fill or the value 0 when the notification is of a miss. When the value of the flicker is positive and the MLO is filled, the liquidity providers that were offering liquidity at the best quote are likely to be adversely selected. Next, approximate $\tilde{H}$ with a $\mathcal{C}^\infty$ function f^ε, where the parameter ε controls the convergence of f^ε to $\tilde{H}$ as $\varepsilon \searrow 0$. For example, to determine the fill or miss outcome of a trade attempt, we use instead of $\tilde{H}(x)$ the sigmoid function evaluated at $-x/\varepsilon$. We recall that the sigmoid function is given by $S(x) = 1/(1 + \exp(-x))$. We work with an approximation to the step function $\tilde{H}$ to preserve the continuity of the impulse operator. **To simplify notation**, we drop the superscript ε and refer to f^ε as f.

3.3 Admissible strategies

The trader controls when to send MLOs and controls the limit rate of the MLOs; this is summarised in the execution strategy $\alpha = (\tau_i, \mathfrak{l}_i)_{i \geq 1}$. The limit rates $(\mathfrak{l}_i)_{i \geq 1}$ of the MLOs are $\mathcal{F}_{\tau_i}$-measurable and take values in a non-empty compact set $L \subseteq \mathbb{R}$. We assume that the trader does not attempt a trade until the outcome of the previous MLO

is known. Thus we require that the $\mathbb{F}$-stopping times in the sequence $(\tau_i)_{i\geq 1}$ satisfy $\tau_{i+1} \geq \tilde{\tau}_i > \tau_i$ for $i \in \mathbb{N}$. Specifically, the trader's set of admissible strategies is

$$\mathcal{A} = \{\alpha = (\tau_i, \ell_i)_{i\geq 1} : \text{for each } i \geq 1, \tau_i \text{ is an } \mathbb{F}\text{-stopping time,} \\ \tau_{i+1} \geq \tilde{\tau}_i > \tau_i, \ell_i \text{ is valued in } L, \text{ and } \ell_i \text{ is } \mathcal{F}_{\tau_i}\text{-measurable}\}.$$

If at the terminal time T, there is an outstanding MLO in the exchange, the order is cancelled. The trader keeps track of pending orders in the exchange with the process $k_t(\alpha)$ (for notational convenience, we use $k(t, \alpha)$), which returns the value 1 if the trader is waiting to be notified of the outcome of a trade attempt, or the value 0 if there is no pending order waiting to be processed by the exchange.

Hence, $k_t(\alpha) = \text{card}\{i \in \mathbb{N} : \tau_i \leq t, \tilde{\tau}_i > t\} \in \{0, 1\}$, which is adapted to $\mathbb{F}$ because for a fixed $t \in [0, T]$ and $i \in \mathbb{N}$, we have that τ_i and $\tilde{\tau}_i$ are $\mathbb{F}$-stopping times. Thus $\{\tau_i \leq t\} \cap \{\tilde{\tau}_i > t\} \in \mathcal{F}_t$.

3.4 System dynamics

In our framework, MLOs have permanent price impact when the order is filled. Note that if the MLO is not filled, the trader is the only market participant who knows that the trade attempt was not successful; thus, missed MLOs do not have price impact. If market participants had access to the information on missed trades, they would learn about buy and sell pressure in the market and would adjust their liquidity-provision and liquidity-taking strategies.

All trade attempts are of size one unit, which we refer to as child orders. The size of the parent order is $\mathfrak{M} > 0$ child orders, which is the initial inventory the trader seeks to liquidate. If the limit rate of the sell MLO to execute a child order is less than or equal to the best bid rate, we assume that there is enough liquidity to fill the MLO without walking the book, in which case the trader is indifferent between sending an IoC or an FoK to the exchange. Thus we assume that trades do not have temporary price impact; it is straightforward to include this price impact in the price dynamics of our model. In Table 3, we see that the majority of liquidity-taking orders do not walk the LOB. Traders specify quantities that can be filled with the volumes displayed at the best bid and best ask rates in the market.

For an execution strategy $\alpha = (\tau_i, \ell_i)_{i\geq 1} \in \mathcal{A}$, the trader monitors the system

$$X_t^\alpha = (S_t^\alpha, Q_t^\alpha, C_t^\alpha),$$

where $(S_t^\alpha)_{t\geq 0}$ is the fundamental best bid rate process, $(Q_t^\alpha)_{t\geq 0}$ is the inventory of the agent, and $(C_t^\alpha)_{t\geq 0}$ is the cash process. The dynamics of the fundamental best bid rate process are

$$S_t^\alpha = S_0 + \int_0^t b(S_u^\alpha)\,\mathrm{d}u + \int_0^t \sigma(S_u^\alpha)\,\mathrm{d}W_u - \kappa \sum_{\tilde{\tau}_i \leq t} f(\ell_i - S^\alpha_{\tilde{\tau}_i -} - \mathsf{F}_{\tilde{\tau}_i}), \tag{3.4}$$

and the cash and the inventory process satisfy, respectively,

$$C_t^\alpha = \sum_{\tilde{\tau}_i \le t} f(\mathfrak{l}_i - S^\alpha_{\tilde{\tau}_i -} - \mathsf{F}_{\tilde{\tau}_i})(S^\alpha_{\tilde{\tau}_i -} + \mathsf{F}_{\tilde{\tau}_i}),$$

$$Q_t^\alpha = \mathfrak{M} - \sum_{\tilde{\tau}_i \le t} f(\mathfrak{l}_i - S^\alpha_{\tilde{\tau}_i -} - \mathsf{F}_{\tilde{\tau}_i}).$$

In the fundamental bid rate in (3.4), the functions $b, \sigma : \mathbb{R} \to \mathbb{R}$ are Lipschitz-continuous. The function b is the drift of the fundamental bid price and σ is the volatility of the innovations. The last term on the right-hand side represents the permanent price impact that filled MLOs have on the fundamental best bid rate, where $\kappa \ge 0$ is the permanent impact parameter of the bid rate. Recall that f is the $\mathcal{C}^\infty$ approximation of the step function that flags when a trade is filled or missed, and that all MLOs sent by the trader are of size one. We assume that the affected fundamental best bid rate $(S_t^\alpha)_{t\ge 0}$ is bounded by the unaffected fundamental best bid rate $(S_t)_{t\ge 0}$ as follows: for any execution strategy $\alpha \in \mathcal{A}$, we have that $\sup_{0\le t\le T} |S_t^\alpha| \le \sup_{0\le t\le T} |S_t| + \kappa N_T$. This assumption is non-restrictive. If the affected fundamental best bid rate is assumed to be always positive, then $S_t^\alpha \le S_t$ for $t \in [0, T]$ implies that $\sup_{0\le t\le T} |S_t^\alpha| \le \sup_{0\le t\le T} |S_t| \le \sup_{0\le t\le T} |S_t| + \kappa N_T$. Alternatively, if b, σ are constant functions, the assumption is also satisfied.

The trader intervenes in the system at time τ_i (i.e., sends a sell MLO), and due to latency, the outcome is known at the notification time $\tilde{\tau}_i > \tau_i$ when the system evolves from $X^\alpha_{\tilde{\tau}_i -}$ to $X^\alpha_{\tilde{\tau}_i} = \Gamma(X_{\tilde{\tau}_i -}, \mathsf{F}_{\tilde{\tau}_i}, \mathfrak{l}_i)$. Here, the function $\Gamma : \mathbb{R}^3 \times \mathbb{R} \times L \to \mathbb{R}^3$ is the impulse operator

$$\Gamma(s, q, c, \mathsf{f}, \ell) = \big(s - \kappa f(\ell - s - \mathsf{f}), q - f(\ell - s - \mathsf{f}), c + (s + \mathsf{f}) f(\ell - s - \mathsf{f})\big), \quad (3.5)$$

which describes how the system (i.e., best bid rate, inventory, cash) changes at notification times. If the MLO sell order is filled: the first argument on the right-hand side of (3.5) shows that the fundamental bid rate decreases by one tick; the second argument shows that the inventory decreases by one unit; and the last argument shows that the amount of cash increases by the fundamental bid rate plus the flicker. If the sell MLO is not filled, the fundamental bid rate is not affected, and the cash and inventory positions do not change.

We denote by $|\cdot|$ the Euclidean norm, and the operator Γ is a continuous function that satisfies

$$\sup_{(y,\ell)\in\mathbb{R}^4\times L} \frac{|\Gamma(y, \ell)|}{1 + |y|} < \infty. \quad (3.6)$$

The initial state of the system is $X_0 = (S_0, \mathfrak{M}, 0)$. Here, S_0 is the initial value of the fundamental bid rate, we recall that $\mathfrak{M}$ is the number of child orders (lots of equal size) that the trader wishes to liquidate, and the initial value of the cash account is

zero. The controlled system X^α is the solution to the SDE

$$X_t^\alpha = X_0 + \int_0^t \boldsymbol{b}(X_u^\alpha)\,\mathrm{d}u + \int_0^t \boldsymbol{\sigma}(X_u^\alpha)\,\mathrm{d}W_u + \sum_{\tilde{\tau}_i \le t} \big(\Gamma(X^\alpha_{\tilde{\tau}_i -}, \mathsf{F}_{\tilde{\tau}_i}, \mathfrak{l}_i) - X^\alpha_{\tilde{\tau}_i -}\big), \tag{3.7}$$

where $\boldsymbol{b} : \mathbb{R}^3 \to \mathbb{R}^3$ and $\boldsymbol{\sigma} : \mathbb{R}^3 \to \mathbb{R}^3$ are given by $\boldsymbol{b}(x_1, x_2, x_3) = (b(x_1), 0, 0)$ and $\boldsymbol{\sigma}(x_1, x_2, x_3) = (\sigma(x_1), 0, 0)$, respectively, and both are Lipschitz-continuous. We now fix a finite horizon $T < \infty$ and an execution strategy $\alpha \in \mathcal{A}$. It follows that

$$\mathbb{E}\Big[\sup_{0\le s\le T} |X_s^\alpha|^2\Big] < \infty, \tag{3.8}$$

a result we employ below to show that the value function of the trader is well defined.

3.5 Liquidity taking with stochastic delay

For all $t \in [0, T)$, the agent has either one or no pending order in the exchange. Thus we define two sets of admissible strategies: (i) admissible strategies with one pending order, and (ii) admissible strategies with no pending order. If at time $t \in [0, T)$, there is one pending order with price limit $\ell \in L$, then the set of admissible strategies is

$$\mathcal{A}_{t,\ell} = \{\alpha = (\tau_i, \mathfrak{l}_i)_{i\ge 1} \in \mathcal{A} : \tau_1 = t, \mathfrak{l}_1 = \ell\},$$

and if there is no pending order, the set of admissible strategies is

$$\mathcal{A}_t = \{\alpha = (\tau_i, \mathfrak{l}_i)_{i\ge 1} \in \mathcal{A} : \tau_1 \ge t\}.$$

For any $(t, x, \mathsf{f}) \in [0, T] \times (\mathbb{R} \times (-\infty, \mathfrak{M}] \times \mathbb{R}) \times \mathbb{R}$ with $x = (s, q, c)$ and for a pending order with limit price $\ell \in L$ and $\alpha \in \mathcal{A}_{t,\ell}$, we denote by $X^{t,x,\mathsf{f},\ell,\alpha}$ the solution to (3.7) for $t \le s \le T$ with initial data $X_t = x$ and $\mathsf{F}_t = \mathsf{f}$. One can drop the dependence on f in $X^{t,x,\ell,\alpha}$ because for any $\mathsf{f}_1, \mathsf{f}_2 \in \mathbb{R}$, we have $\mathsf{F}^{t,\mathsf{f}_1}_{\tilde{t}} = \mathsf{F}^{t,\mathsf{f}_2}_{\tilde{t}}$. That is, we write

$$X_s^{t,x,\ell,\alpha} = x + \int_t^s \boldsymbol{b}(X_u^{t,x,\ell,\alpha})\,\mathrm{d}u + \int_t^s \boldsymbol{\sigma}(X_u^{t,x,\ell,\alpha})\,\mathrm{d}W_u + \sum_{t<\tilde{\tau}_i\le s} \big(\Gamma(X^{t,x,\ell,\alpha}_{\tilde{\tau}_i -}, \mathsf{F}^{t,0}_{\tilde{\tau}_i}, \mathfrak{l}_i) - X^{t,x,\ell,\alpha}_{\tilde{\tau}_i -}\big).$$

Similarly, when there is no pending order at time t, we denote by $X^{t,x,\alpha}$ the solution to (3.7) for $t \le s \le T$ with $X_t = x$, for any $(t, x) \in [0, T] \times (\mathbb{R} \times (-\infty, \mathfrak{M}] \times \mathbb{R})$.

Fix $(t, x) \in [0, T] \times \mathbb{R} \times (-\infty, \mathfrak{M}] \times \mathbb{R}$, $\ell \in L$, $\alpha_1 \in \mathcal{A}_{t,\ell}$ and $\alpha_0 \in \mathcal{A}_t$. It follows from (3.6), Gronwall's lemma, the Burkholder–Davis–Gundy inequality and (3.3)

that there exists a constant $C > 0$ such that

$$\mathbb{E}\Big[\sup_{t\le s\le T}|X_s^{t,x,\ell,\alpha_1}|^2\Big] < C(1+|x|^2),$$
$$\mathbb{E}\Big[\sup_{t\le s\le T}|X_s^{t,x,\alpha_0}|^2\Big] < C(1+|x|^2).$$

The trader evaluates the performance of the execution strategy α with the function

$$\Pi(\alpha) = g(X_T^\alpha) + \int_0^T h(X_s^\alpha)\,\mathrm{d}s + \sum_{\tilde\tau_i\le T}\mathfrak{C}(X_{\tilde\tau_i-}, \mathsf{F}_{\tilde\tau_i}, \mathfrak{l}_i),$$

where $g:\mathbb{R}^3\to\mathbb{R}$, $h:\mathbb{R}^3\to\mathbb{R}$, $\mathfrak{C}:\mathbb{R}^4\to\mathbb{R}$ are

$$\begin{aligned} g(s,q,c) &= c + q\big(s+\zeta - a(q-1)\big)(1-\rho),\\ h(s,q,c) &= -\phi q^2,\\ \mathfrak{C}(s,q,c,\mathsf{f},\ell) &= -\rho(s+\mathsf{f})\,f(\ell - s - \mathsf{f}). \end{aligned} \tag{3.9}$$

Here, $a \ge 0$ is the terminal inventory penalty parameter, and $\phi \ge 0$ is the running inventory penalty parameter.

The function g consists of three terms: (i) c, the cash accumulated by the strategy, (ii) $q(s+\zeta)$, the mark-to-market value of the inventory (net of fees), which includes the expected value $\zeta = \int_{\mathbb{R}} z\nu(\mathrm{d}z)$ of the flicker, and (iii) $aq(q-1)$, the costs of walking the LOB (net of fees) when the terminal inventory is greater than one child order.

The function h represents the urgency of the agent to complete the execution programme. Everything else being equal, the higher the value of $\phi \ge 0$, the quicker will the trader liquidate inventory at the beginning of the trading interval. Each trader decides the urgency of their trade or set of trades. For example, the value of the urgency parameter is arbitrarily high for a trader who requires immediate execution. On the other hand, a patient liquidity taker who has time to retry missed trades will set the value of ϕ close to or at zero.

The function $\mathfrak{C}$ represents transaction costs per unit of inventory traded. These costs are based on the notional traded, which is $(s+\mathsf{f})\,f(\mathfrak{l}-s-\mathsf{f})$ per one unit of inventory, and $\rho\in(0,1)$ is the transaction cost parameter.

The controlled processes $(C_t^\alpha)_{0\le t\le T}$, $(S_t^\alpha)_{0\le t\le T}$ and $(Q_t^\alpha)_{0\le t\le T}$ obey the bounds, which do not depend on the strategy α or $t\in[0,T]$,

$$\begin{aligned} |C_t^\alpha| &\le \sum_{i=1}^{N_T}|Z_i| + N_T\sup_{0\le u\le T}S_u + \kappa N_T^2,\\ |S_t^\alpha| &\le \sup_{0\le u\le T}|S_u| + \kappa N_T,\\ |Q_t^\alpha| &\le \mathfrak{M} + N_T. \end{aligned} \tag{3.10}$$

Furthermore, we see that

$$\sup_{(x,\mathsf{f},\ell)\in\mathbb{R}^3\times\mathbb{R}\times L}\frac{|g(x)|+|h(x)|+|\mathfrak{C}(x,\mathsf{f},\ell)|}{1+|\mathsf{f}|+|x|^2}<\infty,$$

which ensures that the value functions we present below are well defined, i.e., they are finite as a consequence of the inequality above and (3.8) and (3.10).

When there is one pending order in the exchange, the trader's performance criterion and value function are

$$J_1(t,x,\ell,\alpha)=\mathbb{E}\bigg[g(X_T^{t,x,\ell,\alpha})+\int_t^T h(X_s^{t,x,\ell,\alpha})\,\mathrm{d}s+\sum_{\tilde{\tau}_i\le T}\mathfrak{C}(X_{\tilde{\tau}_i-}^{t,x,\ell,\alpha},\mathsf{F}_{\tilde{\tau}_i}^{t,0},\ell_i)\bigg]$$

and

$$v_1(t,x,\ell)=\sup_{\alpha\in\mathcal{A}_{t,\ell}}J_1(t,x,\ell,\alpha),\tag{3.11}$$

respectively, for $(t,x)\in[0,T]\times\mathbb{R}\times(-\infty,\mathfrak{M}]\times\mathbb{R}$, $\ell\in L$, $\alpha\in\mathcal{A}_{t,\ell}$.

Similarly, when the trader does not have an order pending in the exchange, we have

$$J_0(t,x,\alpha)=\mathbb{E}\bigg[g(X_T^{t,x,\alpha})+\int_t^T h(X_s^{t,x,\alpha})\,\mathrm{d}s+\sum_{\tilde{\tau}_i\le T}\mathfrak{C}(X_{\tilde{\tau}_i-}^{t,x,\alpha},\mathsf{F}_{\tilde{\tau}_i}^{t,0},\mathfrak{l}_i)\bigg],$$

for $(t,x)\in[0,T]\times\mathbb{R}\times(-\infty,\mathfrak{M}]\times\mathbb{R}$, $\alpha\in\mathcal{A}_t$, with the corresponding value function

$$v_0(t,x)=\sup_{\alpha\in\mathcal{A}_t}J_0(t,x,\alpha).\tag{3.12}$$

In Appendix A, we show in Theorem A.1 that the trader's value functions satisfy the dynamic programming principle (DPP). The part of the proof that deals with measurable selection arguments is omitted (see Bouchard and Touzi [8] for the derivation of the dynamic programming equation in the sense of viscosity solutions). In Sect. A.1, we derive the HJBQVI satisfied by the value functions, and in Sect. A.2, we study the viscosity properties of the value functions.

3.6 Benchmarks

In Appendix B, we develop two new strategies to benchmark the performance of the random-latency-optimal trading strategy. Throughout, most of the notation is the same as in the previous sections. We derive optimal execution strategies when latency is zero in the marketplace, and when latency is greater than zero and deterministic, respectively. In both cases, the trader solves an impulse control problem. In the first model, the strategy determines the optimal times to send MOs to the exchange, and in the second, the strategy determines the timing and the limit rate of each MLO. We remark that most approaches in the execution literature assume that agents send MOs at a continuous rate. Although agents cannot continuously trade in the market,

this assumption is convenient because in some cases one finds strategies in closed form; see Cartea et al. [12, Sect. 6] and Guéant [20, Part II]. One notable exception is the paper by Cartea and Jaimungal [10], where the authors solve an impulse control problem in which the trader employs both MOs and limit orders to execute a large position in a financial instrument when there is zero latency in the marketplace.

4 Performance of execution strategy

We employ ultra-high-frequency market data for the currency pair EUR/USD to compare the performance of the random-latency-optimal strategy (RLOS) developed in Sect. 3 with four benchmarks (of which the first two are characterised in Appendix B): (i) deterministic-latency-optimal strategy (DLOS), (ii) zero-latency-optimal strategy (ZLOS), (iii) time-weighted-average-price (TWAP), and (iv) execution now (ENOW). (RLOS, DLOS and ZLOS are computed using finite differences. Convergence results for these algorithms remain to be studied.) TWAP sends MOs of equal size and at equally spaced time intervals over the trading window. ENOW is a hypothetical benchmark that assumes there is enough liquidity at the best bid rate in the LOB to execute all the inventory with one MO, with zero latency, at the beginning of the trading window.

Recall that in our formulation, the fundamental bid rate S_t is an input to RLOS, DLOS, ZLOS. However, market participants observe the best bid rate posted on the LOB of the exchange, a rate that consists of the fundamental bid rate S_t and short-lived deviations. Here, we employ LOB observations of the best bid rates and of the quoted spreads in the LOB of the exchange to estimate the fundamental bid rate S_t. In Sect. 2.3, the bottom panel of Fig. 1 depicted the evolution of the best rates in the LOB of the EUR/USD currency pair, where we saw that most flickers cause short-lived widening of the quoted spread: most flickers in the best bid rate are positive, and most flickers in the best ask rate are negative. Thus we assume here that at time t, the fundamental best bid rate S_t is the bid rate in the LOB the last time the spread in the LOB was less than υ ticks. **Throughout**, we assume that $\upsilon = 11$ ticks – our results are robust to various choices of the value of υ.

The trader's execution horizon is $T = 6$ seconds, and the inventory to liquidate is $\mathfrak{M} = 10$ lots, where each lot is €500'000; so the objective is to exchange €5'000'000 into USD. The execution value of ENOW in a frictionless market is the value of $\mathfrak{M}$ lots exchanged at the best bid rate in the LOB at the start of the trading horizon. For TWAP and ZLOS, the MO for each lot is processed with the liquidity posted at the best bid rate in the LOB of the exchange, and if necessary, the order will walk down until filled in full. Similarly, if the limit rate of the MLO allows it, RLOS and DLOS will also walk down the LOB if the liquidity in the best quotes is not enough – we do so even though the strategies characterised in this paper assume that orders of size one do not walk the LOB.

To compute RLOS, DLOS, ZLOS, we assume that the fundamental best bid rate satisfies

$$S_t = S_0 + \sigma W_t, \tag{4.1}$$

where σ is a volatility parameter. However, we remark that when we implement the liquidation strategies, we do not simulate the fundamental best bid rate and do not simulate the flickers that affect the MLO when it is processed by the exchange. Instead, we use the best bid rate in the LOB of LMAX. The only aspect of the model that we simulate is the latency of the trader that employs the strategy – in Sect. 4.5, we perform robustness checks with respect to latency. Thus the results we report are robust to model and parameter misspecification. We assume that our trades do not impact the dynamics of the LOB. Ideally, one would like to test the model when other market participants react to the activity generated by the latency-optimal strategy; however, it is not possible to endogenise the reaction of other market participants to our trades.

We employ data from 1 August 2019 to 31 August 2019 to estimate the parameters of the model, and data between 1 September 2019 and 29 February 2020 to implement and compute the performance of RLOS and the benchmark strategies. Each day, we implement the liquidation strategy over three ten-minute windows. We use the data in the ten minutes starting at 9.00am, at 2.00pm and at 7.00pm for each trading day between 1 September 2019 and 29 February 2020. Times reported in this paper are in London time. At 9.00am and 2.00pm, the foreign exchange activity tends to be very high because by 9.00am, many financial markets in London are open and very active, and many financial markets in the US open at around 2.00pm. Later in the afternoon, foreign exchange activity decreases, and by 7.00pm, many markets worldwide are outside the hours of their main trading activity.

4.1 Model parameters

We employ data from August 2019 to estimate the parameters of the model as follows:

(i) *Volatility of fundamental best bid rate dynamics*: We compute the quadratic variation of the fundamental best bid rate to estimate the volatility parameter σ in (4.1). The estimates over the three ten-minute windows are: (9.00am to 9.10am) $\hat{\sigma}_{9\text{am}} = 2.1 \times 10^{-4}$; (2.00pm to 2.10pm) $\hat{\sigma}_{2\text{pm}} = 2.9 \times 10^{-4}$; (7.00pm to 7.10pm) $\hat{\sigma}_{7\text{pm}} = 1.7 \times 10^{-4}$. The estimate of the volatility from 9.00am to 8.00pm is given by $\hat{\sigma} = 1.9 \times 10^{-4}$; we use this parameter estimate in Sect. 4.4 when trading around news events.

(ii) *Distribution of the size of the flicker that affects the MLOs at processing times*: Recall that the law of the flickers is

$$\nu(\mathrm{d}z) = p_0\delta_{\{0\}}(\mathrm{d}z) + \underbrace{p_+\eta_+ e^{-\eta_+ z}\mathbf{1}_{\{z>0\}}\,\mathrm{d}z}_{\text{price improvement}} + \underbrace{p_-\eta_- e^{\eta_- z}\mathbf{1}_{\{z<0\}}\,\mathrm{d}z}_{\text{slippage}}.$$

Agents use their trade activity to estimate their expected latency in the marketplace; here, we assume that their estimate is one of 10, 30, 60, 90 ms. Next, for a given expected latency and for a value of υ to obtain the fundamental best bid rate, agents estimate the parameters of the distribution of flickers as follows: employ LOB data to study the hit and miss of MLOs with limit rates equal to the fundamental best bid rate when latency is exponentially distributed; use ten minutes of LOB data (beginning at

Table 4 Trading window 9.00am to 9.10am. Fundamental best bid rate is obtained with spread filter of $\upsilon = 11$ ticks. EL stands for expected latency

$\hat{\sigma}_{9\text{am}}$	$\hat{p}_+$	$\hat{p}_-$	$\hat{\eta}_+, \hat{\eta}_-$	$\hat{a}$	EL	ϕ	κ	$\mathfrak{c}$
2.1×10^{-4}	0.05	0.08	$\{1.87, 2.66\} \times 10^{-5}$	1.80×10^{-5}	30 ms	0	0	3×10^{-6}

9.00am, 2.00pm and 7.00pm) of each trading day in August to simulate over one million MLOs; compute the SPIs of the MLOs; and finally, obtain maximum likelihood estimates of the parameters $p_0, p_+, p_-, \eta_+, \eta_-$ for each ten-minute trading interval and for each expected latency.

(iii) *Terminal penalty*: We employ the liquidity-taking trades that walk the LOB to compute the value of the terminal penalty parameter a. For each order, we look at the excess traded quantity above the liquidity posted at the best quotes and perform a linear regression where the explanatory variable is the excess traded quantity, bought or sold, and the dependent variable is the excess rate paid or received in the quote currency for walking the book. The slope of the regression for data in August 2019 is $\hat{a} = 1.80 \times 10^{-5}$, which the agent uses to compute the optimal strategies. On the other hand, to test the performance of the strategy from September 2019 to February 2020, we use $\hat{a} = 1.61 \times 10^{-5}$, which is the in-sample estimate, and we remark that the trader does not know this value.

We assume that the value of the permanent price impact parameter κ is zero because we use LOB data to run the simulations; so we cannot include the permanent impact that the trader's orders would have had on the market. Recall that in the implementations, the MLOs walk the book if necessary (while obeying the price limit), and the strategies incorporate the fees paid to the exchange. Thus we account for all trading frictions except the friction arising from the permanent price impact. The central scenarios we study assume that the liquidity trader is patient; so we set the value of the urgency parameter ϕ to zero. Finally, the transaction cost parameter is $\mathfrak{c} = 3 \times 10^{-6}$, i.e., 3 USD per million USD traded, which are the fees paid by active traders in the foreign exchange market. Table 4 summarises the model parameters for RLOS, DLOS, ZLOS to trade between 9.00am and 9.10am when expected latency is 30 ms. In the interest of space, we do not report the parameters for the other scenarios we study.

4.2 Simulations and performance measure

There are 128 days of market data in the period from 1 September 2019 to 29 February 2020. For each day, we split the ten-minute window into 100 intervals of six seconds. For each interval of six seconds, we combine the one market run of the best bid rate posted on the LOB with 100 simulated runs of the random latency that the trader would face in each trade attempt. Thus the total number of runs for each ten-minute window (9.00am to 9.10am, 2.00pm to 2.10pm, 7.00pm to 7.10pm) is $128 \times 100 \times 100 = 1'280'000$. In each run, the strategy exchanges 5'000'000 EUR into USD, which is 10 times the volume that one normally observes posted at the best bid rate in the exchange, i.e., $\mathfrak{M} = 10$ lots of 500'000 EUR each.

For RLOS, DLOS, ZLOS, we recall that the cost of liquidating terminal inventory with one MO is as follows: if $Q_{T-} = 1$ lot, the order does not walk the LOB. However, if $Q_T \geq 2$, the first lot does not walk the book, and the remaining $Q_T - 1$ lots walk the LOB, so that the USD received from exchanging the final position in EUR is $Q_{T-}(S_T + \zeta - a(Q_{T-} - 1))(1 - \rho)$; see (3.9).

The measure of performance is

$$\frac{\mathbb{E}[\mathrm{TC}_T(\mathrm{RLOS})] - \mathbb{E}[\mathrm{TC}_T(\mathrm{B})]}{V_0} \times 10^6, \tag{4.2}$$

where TC_T is the terminal cash in USD obtained by the strategy, V_0 is the value of the position to exchange at the beginning of the trading window, and B represents one of the four benchmark strategies. The numerator of (4.2) is in USD and the denominator is in EUR; so the units of the performance measure are \$/€M, i.e., USD per million of EUR exchanged. Finally, when the benchmark is ENOW, the performance measure is 'implementation shortfall', which is a widely used benchmark to assess implicit costs and the liquidity of markets. ENOW is seldom achieved because large orders need much more than the liquidity available at the best quotes of the LOB; see Almgren [3] and Cartea et al. [12, Sect. 4]. Generally, one computes implementation shortfall with the midprice. Here, we use the best bid rate for a liquidation problem.

4.3 Results: performance of RLOS and benchmarks

In this subsection, we present the results of the execution strategy for the three ten-minute trading windows that start at 9.00am, 2.00pm and 7.00pm for various levels of the trader's expected latency in the marketplace. Later, in Sect. 4.4, we examine the performance of the strategies around news announcements. The main finding we report is the superior performance of the latency-optimal strategies RLOS and DLOS. Both RLOS and DLOS perform similarly; we do not find evidence at 95% confidence to reject the hypothesis that their mean performance is the same. The outperformance of RLOS over DLOS is positive in most scenarios, but is always less than \$0.5/€M in absolute value. RLOS and DLOS produce similar performances because the optimal limit rates coincide for most of the simulations when rounded to the nearest tick size. The cash value of the outperformance is approximately the same as the fees paid by traders in foreign exchange markets; the value of the outperformance considerably increases during news events. The outperformance stems from the speculative MLOs sent by RLOS and DLOS, which are more valuable when the market is more active.

There are two key insights. First, the performance of MLOs depends on the distribution of the flickers. When activity in the markets increases, the value of MLOs increases because both the probability of receiving a price improvement and the size of the improvements increase – the value of MLOs is highest during news events. On the other hand, an increase in either the size or the probability of slippage during times of heightened activity has little effect on the performance of latency-optimal strategies because the rate limits in the sell MLOs prevent orders from walking down the LOB. This is in contrast with strategies that send MOs (e.g. TWAP and ZLOS) because these orders are exposed to adverse changes in the rates. Second, traders use speed to complete the execution programme with as many MLOs as possible because

Table 5 Values of performance measure (4.2) from 9.00am to 9.10am, $T = 6$ seconds, $\phi = 0$. The revenues from RLOS and from DLOS (not shown) are statistically the same

	Performance 9.00am–9.10am Expected latency in ms			
	10	30	60	90
ZLOS	4.56	3.40	2.74	2.34
TWAP	5.14	3.80	3.09	2.65
ENOW	3.91	2.57	1.86	1.42

they can retry missed speculative MLOs before reaching the end of the trading interval. Thus the lower the latency of the trader, the more opportunities the trader will have to fill speculative MLOs, so the better is the performance of latency-optimal strategies that use MLOs over strategies that employ MOs.

Cartea and Sánchez-Betancourt [15] estimate the range of the stochastic latency of an active trader in the foreign exchange market to be between 10 and 50 ms. Table 5 reports the results of the performance measure (4.2) over the window 9.00am to 9.10am for a patient trader with 10, 30, 60, 90 ms of expected latency. RLOS receives approximately between 5.14 and 2.65 USD more than TWAP for every million EUR exchanged into USD, and RLOS receives between 4.56 and 2.34 USD more than ZLOS per million EUR exchanged, and RLOS outperforms ENOW by between 1.42 and 3.91 USD per million EUR exchanged. Finally, the average USD amount that RLOS receives is statistically the same as that received by DLOS, so the table does not report the results for DLOS. Here, two means are statistically the same if the Student t-test does not reject the hypothesis that the two means are the same with 95% confidence. All outperformances we report are statistically different from zero. For example, in Table 5, the outperformance of RLOS over TWAP when the expected latency is 30 ms is \$3.80/€M. Here, we reject the null hypothesis that the two means are the same with a p-value of 2.15×10^{-5}.

The results in Table 5 show that the value of the outperformance of RLOS (and DLOS) decreases as the value of the expected latency increases – we provide two stylised facts of the strategies to explain the results in the table. First, as expected latency increases, it is more likely that at the end of the trading interval, all strategies will fall short of liquidating the target inventory. In the extreme case where expected latency is arbitrarily large, it is very likely that by the terminal time T, most of the target inventory will remain in euros, i.e., is not exchanged into USD. Thus at time T, the remaining lots of currency pairs will be liquidated with one MO, where the trader pays the costs of walking the book; the costs are the same for all strategies (except for ENOW).

Second, for a fixed trading window and for a fixed liquidation target, faster traders (i.e., with lower expected latency) have on average more opportunities than slow traders to send MLOs to the exchange (recall that in our model, the trader cannot send another trade if her previous order is pending in the exchange). Therefore, compared with slow traders, faster traders send more (and fill more) MLOs that seek a rate improvement, which explains the results we report in the table – we return to this point below. In an extreme case, a patient trader with nearly zero latency will use her superior speed advantage to send a large amount of MLOs that seek price

Table 6 Expected latency is 30 ms, $T = 6$ seconds, so expected number of attempts is 200, and $\phi = 0$. LR: limit rate of MLO. S_t: fundamental best bid rate when the trader decides to send the MLO

	Expected latency is 30 ms, start time 9.00am					
	MLO Attempts	MLO Fills	$Q_{T-} \geq 2$	LR $< S_t$	LR $= S_t$	LR $> S_t$
RLOS	195.07	9.95	0.00	1.00	5.71	188.35
DLOS	197.44	9.64	0.07	1.26	5.01	191.17
ZLOS	10.00	10.00	0.00	10.00	–	–
TWAP	10.00	10.00	0.00	10.00	–	–
ENOW	1.00	1.00	0.00	–	1.00	–

improvements, and due to the flickers in the best bid rate in the LOB, the trader will exchange a large proportion of the initial inventory in EUR into USD with these speculative MLOs. A very fast and very patient trader will not employ her speed in the marketplace to liquidate the target inventory quickly and ahead of time; the trader will use her low latency to fill as many speculative trades as possible to complete the execution target.

Our results show that the slower the trader, the less she benefits from latency-optimal strategies. However, on the other hand, an implication of our findings is that slow traders may profit from investing in co-location, software and other services to increase their speed in the marketplace if the improvement in the performance of the execution strategies outweighs the costs to reduce their latency, which largely depends on the volumes traded.

To gain further insights, we report summary statistics of the behaviour of the strategies from 9.00am to 9.10am when expected latency is 30 ms. Table 6 reports: the average number of trade attempts (there are 200 periods of 30 ms for an execution horizon of six seconds); the average number of filled orders; the proportion of runs for which the strategy sends an MO at time T to liquidate the terminal inventory and pays extra costs from walking the LOB (i.e., average number of runs for which $Q_{T-} \geq 2$); the proportion of the number of attempts for which the limit rate (LR) of the MLO is lower than the fundamental best bid rate S_t (the trader is willing to walk the LOB to complete the trade); the average number of attempts for which LR $= S_t$, and the average number of attempts where LR $> S_t$, which are speculative trades that target a rate improvement (relative to the fundamental exchange rate).

On average, RLOS and DLOS employ a similar number of MLO attempts, of which RLOS fills 9.95 and DLOS fills 9.65 (these figures exclude the MOs sent, if necessary, at time T). Approximately 97% of the attempted MLOs by both strategies are speculative, of which 2.4% (2.3%) sent by RLOS (DLOS) were filled. As discussed above, filled speculative MLOs are the main source of the superior performance of latency-optimal strategies. Within the trading window, the limit rates of the speculative MLOs become less ambitious if the inventory is not on target and the terminal date approaches. However, the latency-optimal strategies cannot guarantee that at time T, the inventory in EUR is drawn to $Q_{T-} < 2$. The table shows that in approximately 0.3% (6.9%) of the runs, RLOS (DLOS) arrives at the end of the trading horizon with two or more lots that are liquidated with one MO. Finally, the

Table 7 Expected latency is 30 ms, $T = 6$ seconds, so expected number of attempts is 200, and $\phi = 10^{-3}$. LR: limit rate of MLO. S_t: fundamental best bid rate when the trader decides to send the MLO

	Expected latency is 30 ms, start time 9.00am, $\phi = 10^{-3}$					
	MLO Attempts	MLO Fills	$Q_{T-} \geq 2$	LR $< S_t$	LR $= S_t$	LR $> S_t$
RLOS	10.20	10.00	0.00	10.20	–	–
DLOS	10.21	10.00	0.00	10.21	–	–
ZLOS	10.00	10.00	0.00	10.00	–	–
TWAP	10.00	10.00	0.00	10.00	–	–
ENOW	1.00	1.00	0.00	–	1.00	–

percentage of runs where RLOS outperforms is 57% for DLOS, 61% for ZLOS, 67% for TWAP and 55% for ENOW.

As Table 6 reports, when the value of the urgency parameter ϕ is zero, RLOS and DLOS miss approximately 95% of attempts with MLOs – the majority of the MLOs seek a price improvement that does not materialise. To examine the role of the urgency parameter, we implement the execution strategies described above for a range of values of ϕ and report a summary of the findings for $\phi = 5 \times 10^{-5}$ and $\phi = 1 \times 10^{-3}$. These two choices represent impatient liquidity takers who have more urgency to liquidate the position than the patient trader with $\phi = 0$.

When $\phi = 1 \times 10^{-4}$, RLOS misses 85% of the MLOs and outperforms TWAP by 1.68 USD per million euro exchanged. Similarly, when $\phi = 1 \times 10^{-3}$, RLOS misses 2% of the MLOs and outperforms TWAP by 0.79 USD per million euro exchanged. The performance of latency-optimal strategies is similar to that of TWAP because the urgency with which RLOS must liquidate inventory precludes the strategy from sending speculative trades, and because the limit rates of the MLOs are set low to increase the probability of a fill at the expense of price protection. In Table 7, we observe that RLOS sends nearly all MLOs with a limit rate below the fundamental best bid rate to ensure a prompt liquidation and does not send speculative MLOs – due to the trader's impatience, in 99.97% of the simulations, the strategy completes the liquidation early. These results are in stark contrast with those reported in Table 6, where the trader was patient and used the entire window to complete the execution programme with as many speculative MLOs as possible, and where the majority of orders sent are retries of missed speculative MLOs.

When we amalgamate the MLOs of all traders in our data set between 1 September 2019 and 29 February 2020 in the EUR/USD currency pair, we find that 57.1% of MLOs sent by all traders missed the trade. On the other hand, when we look at the MLOs sent by each trader, we find that approximately 35% of the traders miss up to 25% of their MLOs, and approximately 6% of traders miss between 75% and 100% of their MLOs – these traders also send other types of orders to the exchange (e.g. MOs), see Table 1. Finally, one cannot tell from the data which trades are part of an execution programme (i.e., child orders) or which trades are a single execution that is not part of a larger order. However, as discussed above, our model is designed to execute one stand-alone order or to execute a large parent order that is split into child orders.

Table 8 Values of performance measure (4.2) from 9.00am to 9.10am (first column), from 2.00pm to 2.10pm (second column), and from 7.00pm to 7.10pm (third column), $T = 6$ seconds, and expected latency is 30 ms

	Performance measure (4.2)		
	9.00am to 9.10am	2.00pm to 2.10pm	7.00pm to 7.10pm
ZLOS	3.40	3.36	1.93
TWAP	3.80	3.51	2.25
ENOW	2.57	2.84	1.17

Table 9 Values of performance measure (4.2) starting at 9.00am, $T = 6$ seconds, and $\phi = 0$

	Trader assumes expected latency of 30 ms True expected latency in marketplace is:		
	20 ms	30 ms	40 ms
ZLOS	3.73	3.39	3.14
TWAP	4.20	3.79	3.50
ENOW	2.97	2.55	2.27

In the remainder of this subsection, we focus on a trader with 30 ms of expected latency. Table 8 reports the performance measure when the ten-minute execution windows start at 9.00am, 2.00pm and 7.00pm. We find that the outperformance of RLOS over ZLOS, TWAP, ENOW is worse when the start time is 7.00pm, which is the least active period we study and one for which the probability of positive flickers in the best bid rate is lowest, so MLOs are less valuable.

Next, we study the performance of RLOS when the trader estimates 30 ms of expected latency in the market, but her latency is one of 20 ms (underestimated speed in the marketplace), 30 ms (correct speed estimate), 40 ms (overestimated speed in the marketplace). The results, reported in Table 9, show that the value of the outperformance of the RLOS and DLOS over the other strategies is slightly higher (lower) when the agent underestimates (overestimates) expected latency. Recall that over the trading window, for a lower (higher) value of the expected latency, the trader has more (fewer) attempts on average to execute her inventory.

4.4 Trading around news events

Throughout the calendar year, many pieces of news are released according to a schedule. Market participants know the timing of the release, but do not know the content. In general, trade activity and volatility of the exchange rates tend to increase around the time of the news release. Here, we employ the EUR/USD news event information provided by FXStreet between 1 September 2019 and 29 February 2020 (see https://www.fxstreet.com/economic-calendar). During this period, there are 117 news events marked as high impact for EUR and for USD, and the time of the release of each event is timestamped with precision up to one minute.

We use the estimate $\hat{\sigma} = 1.9 \times 10^{-4}$, which is the volatility of the fundamental best bid between 9.00am to 8.00pm (see Sect. 4.1), to compute the trading strategies around the time of the arrival of the scheduled news; results are robust to employing volatility estimates computed with the data around news events in August 2019. The execution horizon is $T = 6$ seconds and the agent trades during 1, 3, 5, 7 and

Table 10 Performance measure around news events, $T = 6$ seconds, $\phi = 0$, and expected latency is 30 ms

	News events: Performance measure (4.2) Minutes around news announcements:				
	1 min	3 mins	5 mins	7 mins	10 mins
ZLOS	36.31	15.16	10.62	8.66	7.21
TWAP	29.45	13.35	9.62	8.12	6.92
ENOW	23.86	10.23	7.28	6.06	5.15

Table 11 Period is from 1 September 2019 to 29 February 2020. Expected latency is 30ms

	Parameters of the distribution of flickers (ex-post)				
	$\hat{p}_-$ slippage	$\hat{p}_0$ on target	$\hat{p}_+$ price improvement	$1/\hat{\eta}_-$ slippage	$1/\hat{\eta}_+$ price improvement
9am	0.057	0.901	0.042	2.68×10^{-5}	1.92×10^{-5}
2pm	0.064	0.889	0.047	2.66×10^{-5}	2.04×10^{-5}
7pm	0.054	0.923	0.023	2.81×10^{-5}	2.00×10^{-5}
News 1 min	0.229	0.715	0.056	1.62×10^{-4}	2.52×10^{-5}
News 3 min	0.158	0.792	0.050	1.08×10^{-4}	2.25×10^{-5}
News 5 min	0.131	0.821	0.047	8.92×10^{-5}	2.18×10^{-5}
News 7 min	0.118	0.835	0.047	7.94×10^{-5}	2.14×10^{-5}
News 10 min	0.107	0.844	0.049	6.89×10^{-5}	2.12×10^{-5}

10 minutes around the news event. For example, the ISM manufacturing index was scheduled to be released at 2.00pm on 3 September 2019; so for the one-minute window, we employ market data from 1.59.30pm to 1.59.36pm for the first simulation, data from 1.59.36pm to 1.59.42pm for the second simulation, and so on, until the tenth simulation from 2.00.24pm to 2.00.30pm. For each six-second execution interval, we perform 100 simulations of the random latency faced by the trader. Thus for the one-minute window, the study consists of $117 \times 10 \times 100 = 117'000$ simulations. Table 10 reports the results of the performance measure in (4.2).

The value of the outperformance of RLOS over ZLOS and TWAP is approximately between two and ten times the value reported in Table 5. This significant improvement in the performance of RLOS follows from an increase in the number of speculative MLOs that are filled. Table 11 shows the ex-post parameters of the distribution of flickers faced by the trader with 30 ms of expected latency. It is clear that during news announcements, the value of speculative MLOs is higher because the probability of a price improvement (respectively, slippage) and the size of the positive (respectively, negative) flickers are greater than those of the distribution of flickers for the ten-minute trading windows that start at 9.00am, 2.00pm and 7.00pm.

4.5 Robustness checks

In the results we showed above, the fundamental best bid rate was the current or most recent observed best bid rate when the spread was less than $\upsilon = 11$ ticks; recall that one tick is 10^{-5} USD. Table 12 shows that the outperformance of RLOS over

Table 12 Performance measure from 9.00am to 9.10am, $T = 6$ seconds, and expected latency 30 ms

	Filter with υ ticks to obtain fundamental price				
	7	9	11	13	∞
ZLOS	3.03	3.45	3.61	3.55	3.34
TWAP	3.29	3.71	3.87	3.81	3.60
ENOW	1.89	2.31	2.47	2.41	2.20

the benchmarks is robust for $\upsilon \in \{7, 9, 11, 13, \infty\}$ ticks, where expected latency is 30 ms. Note that when $\upsilon = \infty$ ticks, the fundamental best bid rate is the observed bid rate, i.e., $\hat{S}_t = S_t$.

The value of the outperformance of RLOS and DLOS over other benchmarks peaks for $\upsilon = 11$ ticks. In addition, we note that for the outperformance of RLOS over TWAP, any two consecutive results are statistically the same. For example, the outperformance of RLOS over TWAP for $\upsilon = 9$ and $\upsilon = 11$ are statistically the same. Finally, the performances of RLOS and DLOS are robust to misspecification of the volatility parameter of the fundamental best rate. We ran simulations with volatility estimates $\sigma \in \{1, 2, 3\} \times 10^{-4}$ and expected latency 30 ms. The performance measure are as those reported above plus minus \$0.4/€M.

5 Conclusions

We have solved a general stochastic impulse control problem in which there is a stochastic delay between the action and its outcome. As an application, an investor liquidates a large position in a financial instrument when there is latency in the marketplace. We have derived the optimal strategies for stochastic and for deterministic latencies, and compared their performance with that of three benchmarks, including TWAP. During normal trading hours, we have found that the latency-optimal strategies outperform the benchmarks by an amount similar to that of the fees paid by traders in the foreign exchange market. Around news events, the value of the outperformance increases to between two and ten times the value of the fees. The superior performance of the latency-optimal strategies stems from both the speculative MLOs sent by the trader and the rate protection provided by the MLOs. We have shown that fast and patient traders use their superior speed to execute the target inventory with as many speculative MLOs as possible; they do not use their speed advantage to finish the execution programme early, or to hit the best quotes they observe in the LOB.

In financial applications, interesting research problems include a study of the effect that latency has on the financial performance of electronic trading strategies, e.g. market making, pairs trading and other statistical arbitrage strategies. In particular, one could try to extend the recent works of Kalsi et al. [23] and Cartea et al. [14] to account for latency. In these works, the authors use techniques from rough path theory to solve relevant algorithmic trading problems, where it is assumed that there is no delay between trade attempts and executions. Finally, from a mathematical point of view, a challenging problem is to extend the current framework to include other distributions of the delay in the marketplace, and to extend the model so that there can be more than one pending order in the system when latency is stochastic.

Appendix A: Random latency

To simplify notation, we let $L_0 = \emptyset$ and $L_1 = L$, and we drop the argument of a mathematical object if it is the empty set. For example, we write $\mathcal{A}_t$ instead of $\mathcal{A}_{t,\emptyset}$, or $X^{t,x,\alpha}$ instead of $X^{t,x,\emptyset,\alpha}$, etc. Finally, when the index of a function is the empty set, the function returns the empty set. For $k \in \{0, 1\}$, the domains of the value functions v_k are

$$\mathcal{D}_k = \{(t, x, \ell) : (t, x) \in [0, T) \times \mathbb{R} \times (-\infty, \mathfrak{M}] \times \mathbb{R} \text{ and } \ell \in L_k\}.$$

Here, we prove the DPP satisfied by $v_k(t, x, \ell)$, $k \in \{0, 1\}$, $t \in [0, T]$ and $\ell \in L_k$. We introduce the function $\iota(t, \alpha)$ which returns the index of the pending order (if any) at time t and strategy α. That is,

$$\iota(t, \alpha) = \begin{cases} i & \text{if there is a pending order, i.e., } i \text{ such that } \tau_i \le t \text{ and } \tilde{\tau}_i > t, \\ \emptyset & \text{otherwise.} \end{cases}$$

Theorem A.1 *The value functions in (3.11) and (3.12) satisfy the dynamic programming principle (DPP). That is, for $k \in \{0, 1\}$, $(t, x, \ell) \in \mathcal{D}_k$ and $\theta \in \mathcal{T}_{t,T}$, the set of stopping times valued in $[t, T]$, we have*

$$v_k(t, x, \ell) = \sup_{\alpha \in \mathcal{A}_{t,\ell}} \mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,\ell,\alpha})\, \mathrm{d}s + \sum_{\tilde{\tau}_i \le \theta} \mathfrak{C}(X_{\tilde{\tau}_i -}^{t,x,\ell,\alpha}, \mathsf{F}_{\tilde{\tau}_i}^{t,0}, \mathfrak{l}_i) + v_{k(\theta,\alpha)}(\theta, X_\theta^{t,x,\ell,\alpha}, \mathfrak{l}_{\iota(\theta,\alpha)}) \Bigg],$$

where $\mathfrak{l}_\emptyset = \emptyset$. In other words, for $k \in \{0, 1\}$, $(t, x, \ell) \in \mathcal{D}_k$, we have:

(a) ***DPP1.*** *For all $\alpha \in \mathcal{A}_{t,\ell}$, we have for all stopping times θ valued in $[t, T]$ that*

$$v_k(t, x, \ell) \le \mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,\ell,\alpha})\, \mathrm{d}s + \sum_{\tilde{\tau}_i \le \theta} \mathfrak{C}(X_{\tilde{\tau}_i -}^{t,x,\ell,\alpha}, \mathsf{F}_{\tilde{\tau}_i}^{t,0}, \mathfrak{l}_i) + v_{k(\theta,\alpha)}(\theta, X_\theta^{t,x,\ell,\alpha}, \mathfrak{l}_{\iota(\theta,\alpha)}) \Bigg].$$

(b) ***DPP2.*** *For all $\epsilon > 0$, there exists $\alpha \in \mathcal{A}_{t,\ell}$ such that for all stopping times θ valued in $[t, T]$, we have*

$$v_k(t, x, \ell) \ge \mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,\ell,\alpha})\, \mathrm{d}s + \sum_{\tilde{\tau}_i \le \theta} \mathfrak{C}(X_{\tilde{\tau}_i -}^{t,x,\ell,\alpha}, \mathsf{F}_{\tilde{\tau}_i}^{t,0}, \mathfrak{l}_i) + v_{k(\theta,\alpha)}(\theta, X_\theta^{t,x,\ell,\alpha}, \mathfrak{l}_{\iota(\theta,\alpha)}) \Bigg] - \epsilon.$$

Before proceeding to the proof, we remark that from the dynamics of the controlled system X^α in (3.7), the independence of the marks and the memoryless property of the exponential distribution of latency, we have the following properties:

(1) *Markov property of* $(X^\alpha, k(\cdot,\alpha), \mathfrak{l}_{\iota(\cdot,\alpha)})$: For any $\alpha \in \mathcal{A}$, we have

$$\mathbb{E}[\psi(X^\alpha_{\theta_2})|\mathcal{F}_{\theta_1}] = \mathbb{E}\big[\psi(X^\alpha_{\theta_2})\big|\big(X^\alpha_{\theta_1}, k(\theta_1,\alpha), \mathfrak{l}_{\iota(\theta_2,\alpha)}\big)\big], \tag{A.1}$$

for any bounded measurable function ψ and stopping times $\theta_1 \le \theta_2$ a.s.

(2) *Causality of the control*: For any $\alpha = (\tau_i, \mathfrak{l}_i)_{i\ge1} \in \mathcal{A}$ and stopping time θ,

$$\alpha^\theta \in \mathcal{A}_{\theta,\mathfrak{l}_{\iota(\theta,\alpha)}} \qquad \text{and} \qquad \mathfrak{l}_{\iota(\theta,\alpha)} \in L_{k(\theta,\alpha)}, \tag{A.2}$$

where $\alpha^\theta = (\tau_{i+\iota(\theta,\alpha)}, \mathfrak{l}_{i+\iota(\theta,\alpha)})_{i\ge1}$.

(3) *Pathwise uniqueness of the state process*:

$$X^{t,x,\ell,\alpha} = X^{\theta, X^{t,x,\ell,\alpha}_\theta, \mathfrak{l}_{\iota(\theta,\alpha)}, \alpha^\theta} \tag{A.3}$$

on $[0,T]$, for any $(t,x,\ell) \in \mathcal{D}_k, k = 0,1, \alpha \in \mathcal{A}_{t,\ell}$ and $\theta \in \mathcal{T}_{t,T}$.

Proof of Theorem A.1 (a) Fix $(t,x,\ell) \in \mathcal{D}_k$, $k \in \{0,1\}$ and take an arbitrary control $\alpha \in \mathcal{A}_{t,\ell}$. By the law of iterated conditional expectations, we have

$$\begin{aligned}
&J_k(t,x,\ell,\alpha)\\
&= \mathbb{E}\bigg[\int_t^\theta h(X^{t,x,\ell,\alpha}_s)\,\mathrm{d}s + \sum_{\tilde\tau_i\le\theta} \mathfrak{C}(X^{t,x,\ell,\alpha}_{\tilde\tau_i-}, \mathsf{F}^{t,0}_{\tilde\tau_i}, \mathfrak{l}_i)\\
&\qquad + \mathbb{E}\bigg[\int_\theta^T h(X^{t,x,\ell,\alpha}_s)\,\mathrm{d}s + \sum_{\theta<\tilde\tau_i\le T} \mathfrak{C}(X^{t,x,\ell,\alpha}_{\tilde\tau_i-}, \mathsf{F}^{t,0}_{\tilde\tau_i}, \mathfrak{l}_i) + g(X^{t,x,\ell,\alpha}_T)\Big|\mathcal{F}_\theta\bigg]\bigg],
\end{aligned}$$

and because of the joint Markov property of $(X^\alpha, k(\cdot,\alpha), \mathfrak{l}_{\iota(\cdot,\alpha)})$, the causality of the control and the pathwise uniqueness of $X^{t,x,\ell,\alpha}$ (see (A.1)–(A.3)), we have

$$\begin{aligned}
&J_k(t,x,\ell,\alpha)\\
&= \mathbb{E}\bigg[\int_t^\theta h(X^{t,x,\ell,\alpha}_s)\,\mathrm{d}s + \sum_{\tilde\tau_i\le\theta} \mathfrak{C}(X^{t,x,\ell,\alpha}_{\tilde\tau_i-}, \mathsf{F}^{t,0}_{\tilde\tau_i}, \mathfrak{l}_i)\\
&\qquad + J_{k(\theta,\alpha)}(\theta, X^{t,x,\ell,\alpha}_\theta, \mathfrak{l}_{\iota(\theta,\alpha)}, \alpha^\theta)\bigg]\\
&\le \mathbb{E}\bigg[\int_t^\theta h(X^{t,x,\ell,\alpha}_s)\,\mathrm{d}s + \sum_{\tilde\tau_i\le\theta} \mathfrak{C}(X^{t,x,\ell,\alpha}_{\tilde\tau_i-}, \mathsf{F}^{t,0}_{\tilde\tau_i}, \mathfrak{l}_i) + v_{k(\theta,\alpha)}(\theta, X^{t,x,\ell,\alpha}_\theta, \mathfrak{l}_{\iota(\theta,\alpha)})\bigg].
\end{aligned}$$

Now take the supremum over $\alpha \in \mathcal{A}_{t,\ell}$ on both sides to obtain the desired result.

(b) Fix $(t,x,\ell) \in \mathcal{D}_k$, $k = 0,1$ and take an arbitrary control $\alpha \in \mathcal{A}_{t,\ell}$. By the definition of v_k, for any $\epsilon > 0$ and $\omega \in \Omega$, there exists $\alpha_{\epsilon,\omega} \in \mathcal{A}_{\theta(\omega),\mathfrak{l}_{\iota(\theta(\omega),\alpha(\omega))}}$ which is an ϵ-optimal control for $v_{k(\theta(\omega),\alpha(\omega))}$ at the point

$$\big(\theta(\omega), X^{t,x,\ell,\alpha}_{\theta(\omega)}, \mathfrak{l}_{\iota(\theta(\omega),\alpha(\omega))}\big).$$

Next, by a measurable selection argument, see Bertsekas and Shreve [7, Chap. 7], one can show that there is $\bar{\alpha}_\epsilon \in \mathcal{A}_{\theta,\mathfrak{l}_{\iota(\theta,\alpha)}}$ such that $\bar{\alpha}_\epsilon(\omega) = \alpha_{\epsilon,\omega}(\omega)$ for almost all $w \in \Omega$, so that

$$v_{k(\theta,\alpha)}(\theta, X_\theta^{t,x,\ell,\alpha}, \mathfrak{l}_{\iota(\theta,\alpha)}) - \epsilon \le J_{k(\theta,\alpha)}(\theta, X_\theta^{t,x,\ell,\alpha}, \mathfrak{l}_{\iota(\theta,\alpha)}, \bar{\alpha}_\epsilon). \tag{A.4}$$

Next, concatenate the controls α and $\bar{\alpha}_\epsilon$; that is, take $\bar{\alpha}$ to be α for the impulses that are before (or include) time θ, and take $\bar{\alpha}_\epsilon$ for the impulses that are strictly after time θ. By construction, $\bar{\alpha} \in \mathcal{A}_{t,\ell}$ so that $X_u^{t,x,\ell,\alpha} = X_u^{t,x,\ell,\bar{\alpha}}$ for all $u \in [t,\theta]$, $k(\theta,\alpha) = k(\theta,\bar{\alpha})$, $\mathfrak{l}_{\iota(\theta,\alpha)} = \mathfrak{l}_{\iota(\theta,\bar{\alpha})}$ and $\bar{\alpha}^\theta = \bar{\alpha}_\epsilon$. Therefore, following the same steps as in the proof of (a), we have

$$\begin{aligned} J_k(t,x,\ell,\bar{\alpha}) = \mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,\ell,\bar{\alpha}})\,\mathrm{d}s + \sum_{\tilde{\tau}_i \le \theta} \mathfrak{C}(X_{\tilde{\tau}_i-}^{t,x,\ell,\bar{\alpha}}, \mathsf{F}_{\tilde{\tau}_i}^{t,0}, \mathfrak{l}_i) \\ + J_{k(\theta,\bar{\alpha})}(\theta, X^{t,x,\ell,\bar{\alpha}}, \mathfrak{l}_{\iota(\theta,\bar{\alpha})}, \bar{\alpha}_\epsilon) \Bigg], \end{aligned}$$

which together with (A.4) yields

$$\begin{aligned} J_k(t,x,\ell,\bar{\alpha}) \ge \mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,\ell,\bar{\alpha}})\,\mathrm{d}s + \sum_{\tilde{\tau}_i \le \theta} \mathfrak{C}(X_{\tilde{\tau}_i-}^{t,x,\ell,\bar{\alpha}}, \mathsf{F}_{\tilde{\tau}_i}^{t,0}, \mathfrak{l}_i) \\ + v_{k(\theta,\bar{\alpha})}(\theta, X_\theta^{t,x,\ell,\bar{\alpha}}, \mathfrak{l}_{\iota(\theta,\bar{\alpha})}) \Bigg] - \epsilon, \end{aligned}$$

where $\bar{\alpha}$ can be replaced by α because $\bar{\alpha}$ is less than or equal to θ. Now take the supremum over $\alpha \in \mathcal{A}_{t,\ell}$ on both sides to conclude that

$$\begin{aligned} v_k(t,x,\ell) \ge \sup_{\alpha \in \mathcal{A}_{t,\ell}} \mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,\ell,\alpha})\,\mathrm{d}s + \sum_{\tilde{\tau}_i \le \theta} \mathfrak{C}(X_{\tilde{\tau}_i-}^{t,x,\ell,\alpha}, \mathsf{F}_{\tilde{\tau}_i}^{t,0}, \mathfrak{l}_i) \\ + v_{k(\theta,\alpha)}(\theta, X_\theta^{t,x,\ell,\alpha}, \mathfrak{l}_{\iota(\theta,\alpha)}) \Bigg] - \epsilon, \end{aligned}$$

which completes the proof because ϵ is arbitrary. □

We characterise the value functions with the following two corollaries which are stated without proof (they follow from Theorem A.1). In what follows, Z is a random variable with law $\nu(\mathrm{d}x)$ as in (3.2), independent of W and of the exponentially distributed delays between trade attempts and notifications.

Corollary A.2 *For $k = 1$, $(t, x, \ell) \in \mathcal{D}_1$, $p \in \mathbb{R}$ and $\theta \in \mathcal{T}_{t,T}$, we have*

$$v_1(t,x,\ell) = \mathbb{E}\Bigg[\int_t^{\theta\wedge\tilde{\iota}} h(X_s^{t,x})\,\mathrm{d}s + \Big(v_0\big(\tilde{\iota}, \Gamma(X_{\tilde{\iota}-}^{t,x}, Z, \ell)\big) + \mathfrak{C}(X_{\tilde{\iota}-}^{t,x}, Z, \ell)\Big)\mathbf{1}_{\{\tilde{\iota}\le\theta\}} + v_1(\theta, X_\theta^{t,x}, \ell)\mathbf{1}_{\{\tilde{\iota}>\theta\}}\Bigg].$$

In particular, when $\theta = T$, it is possible to characterise v_1 in terms of v_0 via

$$v_1(t,x,\ell) = \mathbb{E}\Bigg[\int_t^{T\wedge\tilde{\iota}} h(X_s^{t,x})\,\mathrm{d}s + \Big(v_0\big(\tilde{\iota}, \Gamma(X_{\tilde{\iota}-}^{t,x}, Z, \ell)\big) + \mathfrak{C}(X_{\tilde{\iota}-}^{t,x}, Z, \ell)\Big)\mathbf{1}_{\{\tilde{\iota}\le T\}} + v_0(T, X_T^{t,x})\mathbf{1}_{\{\tilde{\iota}>T\}}\Bigg] \tag{A.5}$$

and $v_1(T, x, \ell) = v_0(T, x) = c + q(s + \zeta - a(q-1))(1-\rho)$, for all $\ell \in L$.

Corollary A.3 *For $(t, x) \in \mathcal{D}_0$ and $\theta \in \mathcal{T}_{t,T}$, we have*

$$v_0(t,x) = \sup_{(\tau,\mathfrak{l})\in\mathcal{I}_t} \mathbb{E}\Bigg[\int_t^{\theta\wedge\tau} h(X_s^{t,x})\,\mathrm{d}s + v_0(\theta, X_\theta^{t,x})\mathbf{1}_{\{\theta<\tau\}} + v_1(\tau, X_\tau^{t,x}, \mathfrak{l})\mathbf{1}_{\{\tau\le\theta\}}\Bigg], \tag{A.6}$$

where $\mathcal{I}_t$ is the set of all pairs $(\tau, \mathfrak{l})$ where τ is a stopping time with $t \le \tau < T$ or $\tau = \infty$ a.s., and $\mathfrak{l}$ is an $\mathcal{F}_\tau$-measurable random variable valued in L.

We use the following corollary to prove the subsolution property of the value function (see Theorem A.9).

Corollary A.4 *For $(t, x) \in \mathcal{D}_0$ and $\theta \in \mathcal{T}_{t,T}$, we have*

$$v_0(t,x) = \sup_{(\tau,\mathfrak{l})\in\mathcal{I}_t} \mathbb{E}\Bigg[\int_t^{\theta\wedge\tilde{\tau}} h(X_s^{t,x})\,\mathrm{d}s + v_0(\theta, X_\theta^{t,x})\mathbf{1}_{\{\theta<\tau\}} + v_1(\theta, X_\theta^{t,x}, \mathfrak{l})\mathbf{1}_{\{\tau\le\theta<\tilde{\tau}\}} + v_0\big(\tilde{\tau}, \Gamma(X_{\tilde{\tau}-}^{t,x}, Z, \mathfrak{l})\big)\mathbf{1}_{\{\tau<\tilde{\tau}<\theta\}}\Bigg].$$

A.1 PDE system viscosity characterisation

In Corollaries A.2 and A.3, we let $\theta \to t$ to characterise the value functions v_0 and v_1.

Theorem A.5 *The value functions satisfy the dynamic programming equation*

$$\min\left\{-\frac{\partial v_0}{\partial t}(t,x)-\mathcal{L}v_0(t,x)-h(x),\, v_0(t,x)-\sup_{\ell\in L} v_1(t,x,\ell)\right\}=0 \qquad \text{on } \mathcal{D}_0 \tag{A.7}$$

and

$$\begin{aligned}&\frac{\partial v_1}{\partial t}(t,x,\ell)+\mathcal{L}v_1(t,x,\ell)+h(x)-\lambda v_1(t,x,\ell)\\&+\lambda\mathbb{E}\big[v_0\big(t,\Gamma(x,Z,\ell)\big)+\mathfrak{C}(x,Z,\ell)\big]=0 \qquad \text{on } \mathcal{D}_1,\end{aligned} \tag{A.8}$$

where $\lambda>0$ is the parameter of the exponential distribution that describes the random latency,

$$\mathcal{L}=b(s)\frac{\partial}{\partial s}+\frac{1}{2}\sigma^2(s)\frac{\partial^2}{\partial s^2},$$

the random variable Z has law $\nu(\mathrm{d}x)$ as in (3.2) and the terminal conditions are

$$v_1(T,x,\ell)=v_0(T,x)=c+q\big(s+\zeta-a(q-1)\big)(1-\rho) \qquad \text{for all } \ell\in L.$$

The intuition for the result in Theorem A.5 is as follows. With no pending orders, the trader will submit an order with limit rate ℓ^* when it is better to have one order pending in the exchange than none; this decision is determined by the QVI in (A.7). The crucial tradeoff is that once the agent sends an order, she must wait until it is processed by the exchange before sending another order; thus, a pending order precludes the trader from taking advantage of any opportunities that arise while a pending order is being processed during the period of latency. Finally, the trader's value function satisfies the dynamic programming equation in (A.8) when there is an order pending in the exchange.

A priori, in (A.7), it is not clear that the candidate optimal control is well defined. If $v_1(t,x,\ell)$ is continuous in ℓ, then given that L is compact, we can define the candidate optimal control as $\ell^*=\operatorname{argmax}_\ell v_1(t,x,\ell)$ whenever $v_0(t,x)=v_1(t,x,\ell^*)$. On the other hand, if $v_1(t,x,\ell)$ is not continuous in ℓ, we can take the set L to be a finite collection of points in $\mathbb{R}$. Recall that L is the set of possible limit rates, and limit rates in EUR/USD take values with precision of (at most) five decimals. Thus we can define L to be all limit rates with five decimals of precision between zero and a fixed $\bar{\ell}$ – in practice, any number that is twice the current level of the exchange rate suffices.

We use (A.5) to make a key simplification to the system in (A.7) and (A.8). We plug (A.5) into (A.7) to obtain an HJBQVI that characterises the value function v_0. We remark that once we know v_0, we compute v_1 by means of (A.5). Equation (A.7)

becomes

$$\min\Bigg\{-\frac{\partial v_0}{\partial t}(t,x)-\mathcal{L}v_0(t,x)-h(x),\, v_0(t,x)$$
$$-\sup_{\ell\in L}\mathbb{E}\Bigg[\int_t^{T\wedge\tilde{\iota}} h(X_s^{t,x})\,\mathrm{d}s+\Big(v_0\big(\tilde{\iota},\Gamma(X_{\tilde{\iota}-}^{t,x},Z,\ell)\big)+\mathfrak{C}(X_{\tilde{\iota}-}^{t,x},Z,\ell)\Big)\mathbf{1}_{\{\tilde{\iota}\le T\}}$$
$$+v_0(T,X_T^{t,x})\mathbf{1}_{\{\tilde{\iota}>T\}}\Bigg]\Bigg\}=0\qquad\text{on }\mathcal{D}_0,\tag{A.9}$$

with terminal condition $v_0(T,x)=c+q(s+\zeta-a(q-1))(1-\rho)$. We propose the ansatz $v_0(t,s,q,c)=c+h_0(t,s,q)$ to reduce the system (A.9) to

$$0=\min\Bigg\{-\frac{\partial h_0}{\partial t}(t,s,q)-\mathcal{L}h_0(t,s,q)+\phi q^2,$$
$$h_0(t,s,q)+q\Big(q\frac{\phi}{\lambda}(1-e^{-\lambda(T-t)})+a(q-1)e^{-\lambda(T-t)}\Big)$$
$$-qe^{-\lambda(T-t)}\mathbb{E}[S_T^{t,s}]$$
$$-\sup_{\ell\in L}\mathbb{E}\Big[\Big((1-\rho)(S_{\tilde{\iota}-}^{t,s}+Z)f(\ell-S_{\tilde{\iota}-}^{t,s}-Z)$$
$$+h_0\big(\tilde{\iota},\tilde{\Gamma}(S_{\tilde{\iota}-}^{t,s},Z,q,\ell)\big)\Big)\mathbf{1}_{\{\tilde{\iota}\le T\}}\Big]\Bigg\},\tag{A.10}$$

with terminal condition $h_0(T,s,q)=(qs+q\zeta-aq(q-1))(1-\rho)$, and

$$\tilde{\Gamma}(s,Z,q,\ell)=\big(s-\kappa f(\ell-s-Z),q-f(\ell-s-Z)\big).$$

A.2 Viscosity properties of the value function v_0

The value functions v_0 and v_1 do not need to be smooth a priori. We focus on the viscosity properties of the HJBQVI satisfied by v_0 because we fully characterised v_1 in terms of v_0 (see Corollary A.2). Bruder and Pham [9] use backward and forward iterations on the domains and value functions because they have more than two coupled value functions. In our case, a more 'standard' approach suffices because of the reduction in (A.9). For classical references on the subject of viscosity solutions we refer the reader to Crandall et al. [16, Sects. 1–3] and Fleming and Soner [17, Sect. 5].

For a locally bounded function w on $\mathcal{D}_0$, we denote by $\underline{w}$ (respectively, $\overline{w}$) the lower-semicontinuous (respectively, upper-semicontinuous) envelope of w, given by

$$\underline{w}(t,x)=\liminf_{(t',x')\to(t,x)} w(t',x'),$$
$$\overline{w}(t,x)=\limsup_{(t',x')\to(t,x)} w(t',x'),\qquad (t,x)\in\mathcal{D}_0.$$

For any locally bounded function u on $\mathcal{D}_0$, define the locally bounded function $\mathcal{H}u$ on $\mathcal{D}_0$ by

$$\mathcal{H}u(t,x) = \sup_{\ell\in L}\mathbb{E}\bigg[\int_t^{T\wedge\tilde{t}} h(X_s^{t,x})\,\mathrm{d}s + \Big(u\big(\tilde{t},\Gamma(X_{\tilde{t}-}^{t,x},Z,\ell)\big) + \mathfrak{C}(X_{\tilde{t}-}^{t,x},\ell)\Big)\mathbf{1}_{\{\tilde{t}\le T\}} + u(T,X_T^{t,x})\mathbf{1}_{\{\tilde{t}>T\}}\bigg].$$

Lemma A.6 *Let w be a locally bounded function on $\mathcal{D}_0$. Then $\mathcal{H}\overline{w}$ is upper-semicontinuous, and $\overline{\mathcal{H}w} \le \mathcal{H}\overline{w}$.*

Proof First, we prove that $\mathcal{H}\overline{w}$ is upper-semicontinuous. Fix $(t,x)\in\mathcal{D}_0$ and let $(t_n,x_n)_{n\ge1}$ be a sequence in $\mathcal{D}_0$ converging to (t,x) as $n\to\infty$. There exists a sequence $(\mathfrak{l}_n)_{n\in\mathbb{N}}$ valued in L such that $\mathcal{H}\overline{w}(t_n,x_n)$ is equal to

$$\mathbb{E}\bigg[\int_{t_n}^{T\wedge\tilde{t}_n} h(X_s^{t_n,x_n})\,\mathrm{d}s + \Big(\overline{w}\big(\tilde{t}_n,\Gamma(X_{\tilde{t}_n-}^{t_n,x_n},Z,\mathfrak{l}_n)\big) + \mathfrak{C}(X_{\tilde{t}_n-}^{t_n,x_n},Z,\mathfrak{l}_n)\Big)\mathbf{1}_{\{\tilde{t}_n\le T\}} + \overline{w}(T,X_T^{t_n,x_n})\mathbf{1}_{\{\tilde{t}_n>T\}}\bigg]$$

for $n\ge1$ because $\overline{w}$ is upper-semicontinuous and L is compact. By the Bolzano–Weierstrass theorem, the sequence $(\mathfrak{l}_n)_{n\in\mathbb{N}}$ converges (up to a subsequence) to an element $\check{\mathfrak{l}}\in L$. Thus we get

$$\begin{aligned}
\mathcal{H}\overline{w}(t,x) &\ge \mathbb{E}\bigg[\int_t^{T\wedge\tilde{t}} h(X_s^{t,x})\,\mathrm{d}s + \Big(\overline{w}\big(\tilde{t},\Gamma(X_{\tilde{t}-}^{t,x},Z,\check{\mathfrak{l}})\big) + \mathfrak{C}(X_{\tilde{t}-}^{t,x},Z,\check{\mathfrak{l}})\Big)\mathbf{1}_{\{\tilde{t}\le T\}} + \overline{w}(T,X_T^{t,x})\mathbf{1}_{\{\tilde{t}>T\}}\bigg]\\
&\ge \limsup_{n\to\infty}\mathbb{E}\bigg[\int_{t_n}^{T\wedge\tilde{t}_n} h(X_s^{t_n,x_n})\,\mathrm{d}s + \Big(\overline{w}\big(\tilde{t}_n,\Gamma(X_{\tilde{t}_n-}^{t_n,x_n},Z,\mathfrak{l}_n)\big) + \mathfrak{C}(X_{\tilde{t}_n-}^{t_n,x_n},Z,\mathfrak{l}_n)\Big)\mathbf{1}_{\{\tilde{t}_n\le T\}} + \overline{w}(T,X_T^{t_n,x_n})\mathbf{1}_{\{\tilde{t}_n>T\}}\bigg]\\
&= \limsup_{n\to\infty}\mathcal{H}\overline{w}(t_n,x_n),
\end{aligned}$$

where the first inequality follows because $\check{\mathfrak{l}}\in L$. The second inequality follows because $\bar{w}$ is upper-semicontinuous, Γ is continuous and by Fatou's lemma. Finally, the last equality follows by definition. Hence $\mathcal{H}\overline{w}$ is upper-semicontinuous, as required.

Next, to prove that $\overline{\mathcal{H}w}\le\mathcal{H}\overline{w}$, fix $(t,x)\in\mathcal{D}_0$ and let $(t_n,x_n)_{n\ge1}$ be a sequence in $\mathcal{D}_0$ converging to (t,x) as $n\to\infty$ such that $\mathcal{H}w(t_n,x_n)\to\overline{\mathcal{H}w}(t,x)$. Then by the

properties of the sequence and the upper-semicontinuity of $\overline{w}$, we have that

$$\overline{\mathcal{H}w}(t,x) = \lim_{n\to\infty} \mathcal{H}w(t_n,x_n) \leq \limsup_{n\to\infty} \mathcal{H}\overline{w}(t_n,x_n) \leq \mathcal{H}\overline{w}(t,x),$$

where the last inequality follows from the upper-semicontinuity of $\mathcal{H}\overline{w}$. Thus we obtain $\overline{\mathcal{H}w} \leq \mathcal{H}\overline{w}$. □

Definition A.7 A locally bounded function w on $\mathcal{D}_0$ is a *viscosity supersolution* of (A.9) on $\mathcal{D}_0$ if for all $(t_0,x_0,\psi) \in \mathcal{D}_0 \times \mathcal{C}^2(\mathcal{D}_0)$ such that $\underline{w} - \psi$ attains a minimum at (t_0,x_0), we have

$$\min\Bigg\{ -\frac{\partial\psi}{\partial t}(t_0,x_0) - \mathcal{L}\psi(t_0,x_0) - h(x_0),$$
$$\underline{w}(t_0,x_0) - \sup_{\ell\in L}\mathbb{E}\Bigg[\Big(\underline{w}\big(\tilde{t}_0,\Gamma(X^{t_0,x_0}_{\tilde{t}_0-},Z,\ell)\big) + \mathfrak{C}(X^{t_0,x_0}_{\tilde{t}_0-},Z,\ell)\Big)\mathbf{1}_{\{\tilde{t}_0\leq T\}}$$
$$+ \underline{w}(T,X^{t_0,x_0}_T)\mathbf{1}_{\{\tilde{t}_0>T\}} - \int_{t_0}^{T\wedge\tilde{t}_0} h(X^{t_0,x_0}_s)\,\mathrm{d}s\Bigg]\Bigg\} \geq 0.$$

It is a *viscosity subsolution* of (A.9) on $\mathcal{D}_0$ if for all $(t_0,x_0,\psi) \in \mathcal{D}_0 \times \mathcal{C}^2(\mathcal{D}_0)$ such that $\overline{w} - \psi$ attains a maximum at (t_0,x_0), we have

$$\min\Bigg\{ -\frac{\partial\psi}{\partial t}(t_0,x_0) - \mathcal{L}\psi(t_0,x_0) - h(x_0),$$
$$\overline{w}(t_0,x_0) - \sup_{\ell\in L}\mathbb{E}\Bigg[\Big(\overline{w}\big(\tilde{t}_0,\Gamma(X^{t_0,x_0}_{\tilde{t}_0-},Z,\ell)\big) + \mathfrak{C}(X^{t_0,x_0}_{\tilde{t}_0-},Z,\ell)\Big)\mathbf{1}_{\{\tilde{t}_0\leq T\}}$$
$$+ \overline{w}(T,X^{t_0,x_0}_T)\mathbf{1}_{\{\tilde{t}_0>T\}} - \int_{t_0}^{T\wedge\tilde{t}_0} h(X^{t_0,x_0}_s)\,\mathrm{d}s\Bigg]\Bigg\} \leq 0.$$

Finally, we say that w is a *viscosity solution* of (A.9) on $\mathcal{D}_0$ if it is a viscosity supersolution and a viscosity subsolution of (A.9) on $\mathcal{D}_0$.

Before we state the main theorem of this section, we present a comparison result that we need in the proof of uniqueness of the viscosity solution. The following growth condition for the value function is assumed **for the remainder of the paper**:

$$\sup_{(t,x)\in\mathcal{D}_0} \frac{|v_0(t,x)|}{1+|x|} < \infty. \tag{A.11}$$

This assumption is not too restrictive. If $b(s) \equiv b \in \mathbb{R}$ and $\sigma(s) \equiv \sigma \in \mathbb{R}_+$ in the dynamics of the fundamental best bid rate process, then a calculation shows that the bound is satisfied. Note that in Sect. 4, we use $b(s) \equiv 0$ and $\sigma(s) \equiv \sigma > 0$. The bound in (A.11) is required for the comparison result which guarantees uniqueness. However, if we relax this assumption, it is not clear we could ensure uniqueness.

Proposition A.8 *Let u_0 (respectively, w_0) be a viscosity subsolution (respectively, supersolution) of (A.9) satisfying the growth condition (A.11). Suppose that w_0 satisfies $w_0 \geq \mathcal{H}w_0$ on $\mathcal{D}_0$, and that $\overline{u}_0(T,x) \leq \underline{w}_0(T,x)$ for all $x \in \mathbb{R} \times (-\infty, \mathfrak{M}] \times \mathbb{R}$. Then $\overline{u}_0 \leq \underline{w}_0$ on $\mathcal{D}_0$.*

We state the steps one needs to follow (together with the assumptions needed) to complete the proof. First, we show that for any $\eta > 0$, there is a viscosity η-strict supersolution[1] w_0^η of (A.9) satisfying $w_0^\eta \geq \mathcal{H}w_0^\eta$ on $\mathcal{D}_0$ and such that for $(t,x) \in \mathcal{D}_0$, we have $w_0 + \eta C_1 |x|^2 \leq w_0^\eta \leq w_0 + \eta C_2(1+|x|^2)$ for some positive constants C_1, C_2 independent of η. Such a construction is in Bruder and Pham [9, Lemma 7.4] where we take $m = 1$. It then follows by the linear growth condition that $\eta C_1 |x|^2 - C_2 \leq w_0^\eta$ on $\mathcal{D}_0$ for some positive constants C_1, C_2. Similarly to [9], the main step in the proof of Proposition A.8 is to compare a viscosity subsolution to a viscosity η-strict supersolution. It is then standard to write the sub viscosity inequalities[2] and η-strict supersolution inequality in terms of semi-jets, and then conclude with Ishii's lemma. For the proof, we require (i) continuity of f, (ii) the Lipschitz property of b, σ, (iii) the growth condition (A.11), and (iv) the two inequalities above. Finally, comparison of η-strict supersolutions implies comparison for supersolutions by letting $\eta \to 0$.

Theorem A.9 *The value function v_0 is the unique viscosity solution of the HJBQVI in (A.9) on $\mathcal{D}_0$, satisfying (A.11), with $v_0(T,x) = c + q(s + \zeta - a(q-1))(1-\rho)$ and satisfying $v_0 \geq \mathcal{H}v_0$ on $\mathcal{D}_0$.*

Proof (**Local boundedness**) Let $(t,x) \in \mathcal{D}_0$. From (A.11), there exists $r_1 > 0$ that does not depend on (t,x) such that $|v_0(t,x)| \leq r_1(1+|x|)$. Thus the value function v_0 is locally bounded.

(**Supersolution property**) Let $(t_0, x_0, \psi) \in \mathcal{D}_0 \times \mathcal{C}^2(\mathcal{D}_0)$ be such that $\underline{v_0} - \psi$ attains a minimum at (t_0, x_0) with $\underline{v_0}(t_0, x_0) = \psi(t_0, x_0)$. By the definition of $\underline{v_0}$, there is a sequence $(t_n, x_n)_{n \geq 1} \in \mathcal{D}_0$ such that $v_0(t_n, x_n) \to \underline{v_0}(t_0, x_0)$ with $(t_n, x_n) \to (t_0, x_0)$. Let $\ell_1 \in L$ and $(t,x) \in \mathcal{D}_0$. Use (A.6) and Corollary A.2 with an immediate impulse and price limit ℓ_1, i.e., $\tau = t$, $\theta = T$ and $\mathfrak{l} = \ell_1$, to write

$$
\begin{aligned}
v_0(t,x) &\geq v_1(t,x,\ell_1) \\
&= \mathbb{E}\Bigg[\int_t^{T \wedge \tilde{t}} h(X_s^{t,x})\, \mathrm{d}s + \Big(v_0\big(\tilde{t}, \Gamma(X_{\tilde{t}}^{t,x}, Z, \ell_1)\big) + \mathfrak{C}(X_{\tilde{t}-}^{t,x}, Z, \ell_1) \Big) \mathbf{1}_{\{\tilde{t} \leq T\}} \\
&\qquad + v_0(T, X_T^{t,x}) \mathbf{1}_{\{\tilde{t} > T\}} \Bigg].
\end{aligned}
$$

[1] See Bruder and Pham [9] for a definition of a viscosity η-strict supersolution.

[2] These inequalities would be the analogues of [9, Eqs. (7.22) and (7.23)].

Thus

$$\begin{aligned}
&\liminf_{n\to\infty} v_0(t_n, x_n) \\
&\geq \liminf_{n\to\infty} \mathbb{E}\Bigg[\int_{t_n}^{T\wedge\tilde{t}_n} h(X_s^{t_n,x_n})\,\mathrm{d}s \\
&\qquad + \Big(v_0\big(\tilde{t}_n, \Gamma(X_{\tilde{t}_n}^{t_n,x_n}, Z, \ell_1)\big) + \mathfrak{C}(X_{\tilde{t}_n-}^{t_n,x_n}, Z, \ell_1) \Big) \mathbf{1}_{\{\tilde{t}_n \leq T\}} \\
&\qquad + v_0(T, X_T^{t_n,x_n}) \mathbf{1}_{\{\tilde{t}_n > T\}} \Bigg].
\end{aligned}$$

Next, use Fatou's lemma, the continuity of Γ and the definition of $\underline{v}_0$ to obtain

$$\begin{aligned}
\underline{v_0}(t_0, x_0) \geq \mathbb{E}\Bigg[&\int_{t_0}^{T\wedge\tilde{t}_0} h(X_s^{t,x})\,\mathrm{d}s \\
&+ \Big(\underline{v_0}(\tilde{t}_0, \Gamma(X_{\tilde{t}_0}^{t,x}, Z, \ell_1)) + \mathfrak{C}(X_{\tilde{t}_0-}^{t,x}, Z, \ell_1) \Big) \mathbf{1}_{\{\tilde{t}_0 \leq T\}} \\
&+ \underline{v_0}(T, X_T^{t,x}) \mathbf{1}_{\{\tilde{t}_0 > T\}} \Bigg].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\underline{v_0}(t_0, x_0) \geq \sup_{\ell\in L} \mathbb{E}\Bigg[&\int_{t_0}^{T\wedge\tilde{t}_0} h(X_s^{t,x})\,\mathrm{d}s \\
&+ \Big(\underline{v_0}\big(\tilde{t}_0, \Gamma(X_{\tilde{t}_0}^{t,x}, Z, \ell)\big) + \mathfrak{C}(X_{\tilde{t}_0-}^{t,x}, Z, \ell) \Big) \mathbf{1}_{\{\tilde{t}_0 \leq T\}} \\
&+ \underline{v_0}(T, X_T^{t,x}) \mathbf{1}_{\{\tilde{t}_0 > T\}} \Bigg]
\end{aligned}$$

because the price limit ℓ_1 is arbitrary. Finally, to complete the viscosity supersolution property, use Corollary A.3 with $\tau = \infty$, Itô's formula for $\psi \in \mathcal{C}^2(\mathcal{D}_0)$, the property that $\underline{v}_0 - \psi$ attains a minimum at (t_0, x_0) with $\underline{v}_0(t_0, x_0) = \psi(t_0, x_0)$, and let $\theta \to t_0$ to obtain

$$-\frac{\partial\psi}{\partial t}(t_0, x_0) - \mathcal{L}\psi(t_0, x_0) - h(x_0) \geq 0.$$

We base the next part on Bruder and Pham [9].

(**Subsolution property**) Let $(t_0, x_0, \psi) \in \mathcal{D}_0 \times \mathcal{C}^2(\mathcal{D}_0)$ be such that $\overline{v_0} - \psi$ attains a maximum at (t_0, x_0), $\psi(t_0, x_0) = \overline{v_0}(t_0, x_0)$ and $\psi \geq \overline{v_0}$ on $\mathcal{D}_0$. If we have $\overline{v_0}(t_0, x_0) \leq \mathcal{H}\overline{v_0}(t_0, x_0)$, the desired inequality follows and the proof is complete. On the other hand, if $\overline{v_0}(t_0, x_0) > \mathcal{H}\overline{v_0}(t_0, x_0)$, we proceed by contradiction, i.e., we assume there is $\eta > 0$ such that

$$\eta = -\frac{\partial\psi}{\partial t}(t_0, x_0) - \mathcal{L}\psi(t_0, x_0) - h(x_0) > 0.$$

By the definition of $\overline{v_0}$ and the continuity of ψ, there exist $\tilde{\epsilon} > 0$ and a sequence

$$(t_n, x_n)_{n\geq 1} \subseteq \big((t_0 - \tilde{\epsilon}, t_0 + \tilde{\epsilon}) \times B(x_0, \tilde{\epsilon})\big) \cap \mathcal{D}_0$$

such that

$$v_0(t_n, x_n) \longrightarrow \overline{v_0}(t_0, x_0) \qquad \text{with } (t_n, x_n) \to (t_0, x_0), \tag{A.12}$$

$$-\frac{\partial \psi}{\partial t} - \mathcal{L}\psi - h > \frac{\eta}{2} \qquad \text{on } \big((t_0 - \tilde{\epsilon}, t_0 + \tilde{\epsilon}) \times B(x_0, \tilde{\epsilon})\big) \cap \mathcal{D}_0, \tag{A.13}$$

$$\overline{v_0} - \psi < \frac{\eta}{4} \qquad \text{on } \big((t_0 - \tilde{\epsilon}, t_0 + \tilde{\epsilon}) \times B(x_0, \tilde{\epsilon})\big) \cap \mathcal{D}_0.$$

Here $B(x_0, \tilde{\epsilon})$ is the open ball with centre in x_0 and radius $\tilde{\epsilon}$ under the Euclidean metric. By continuity of ψ and (A.12), we also have

$$\gamma_n := v_0(t_n, x_n) - \psi(t_n, x_n) \longrightarrow 0 \qquad \text{as } n \to \infty. \tag{A.14}$$

By Corollary A.4, for each $n \geq 1$, there is a control $(\tau_n, \iota_n) \in \mathcal{I}_{t_n}$ such that

$$\begin{aligned} &v_0(t_n, x_n) - \frac{\eta}{4}\epsilon_n \\ &\leq \mathbb{E}\bigg[\int_t^{\theta_n \wedge \tilde{\tau}_n} h(X_s^{t,x})\,\mathrm{d}s + v_0(\theta_n, X_{\theta_n}^{t,x})\mathbf{1}_{\{\theta_n < \tau_n\}} \\ &\qquad + v_1(\theta_n, X_{\theta_n}^{t,x}, \iota_n)\mathbf{1}_{\{\tau_n \leq \theta_n < \tilde{\tau}_n\}} + v_0\big(\tilde{\tau}_n, \Gamma(X_{\tilde{\tau}_n -}^{t,x}, Z, \iota_n)\big)\mathbf{1}_{\{\tau_n < \tilde{\tau}_n < \theta_n\}}\bigg]. \end{aligned} \tag{A.15}$$

Choose $\theta_n = (t_n + \epsilon_n) \wedge \vartheta_n \wedge \tilde{t}_n$ with

$$\vartheta_n = \inf\{s \geq t_n : X_s^{t_n, x_n} \notin B(x_0, \tilde{\epsilon}/2), s \leq t_n + \tilde{\epsilon}/2, s \leq T\} \in \mathcal{T}_{t_n, T}$$

and $(\epsilon_n)_{n\in\mathbb{N}}$ a strictly positive sequence that satisfies $\epsilon_n \to 0$ and $\gamma_n/\epsilon_n \to 0$ as $n \to \infty$. After simple inspection of (A.15) and because $\tilde{t}_n$ is the first time there is a jump in N after t_n, write

$$\begin{aligned} &v_0(t_n, x_n) - \frac{\eta}{4}\epsilon_n \\ &\leq \mathbb{E}\bigg[\int_t^{\theta_n} h(X_s^{t,x})\,\mathrm{d}s + v_0(\theta_n, X_{\theta_n}^{t,x})\mathbf{1}_{\{\theta_n < \tau_n\}} + v_1(\theta_n, X_{\theta_n}^{t,x}, \iota_n)\mathbf{1}_{\{\tau_n \leq \theta_n\}}\bigg]. \end{aligned} \tag{A.16}$$

Use Lemma A.6 to write $\overline{\mathcal{H}v_0}(t_0, x_0) \leq \mathcal{H}\overline{v_0}(t_0, x_0) < \overline{v_0}(t_0, x_0) = \psi(t_0, x_0)$. As $\overline{\mathcal{H}v_0}$ is upper semicontinuous, ψ is continuous and $\overline{\mathcal{H}v_0}(t_0, x_0) < \psi(t_0, x_0)$, the inequality $\mathcal{H}v_0 \leq \psi$ holds in a neighbourhood of (t_0, x_0). Thus for n sufficiently large and because $v_1(t_n, x_n, \iota_n) \leq \mathcal{H}v_0(t_n, x_n)$, we obtain

$$v_1(\theta_n, X_{\theta_n}^{t_n, x_n}, \iota_n)\mathbf{1}_{\{\tau_n \leq \theta_n\}} \leq \psi(\theta_n, X_{\theta_n}^{t_n, x_n})\mathbf{1}_{\{\tau_n \leq \theta_n\}}. \tag{A.17}$$

Equation (A.16) together with (A.14) and (A.17) yields

$$\psi(t_n, x_n) + \gamma_n - \frac{\eta}{4}\epsilon_n \leq \mathbb{E}\left[\int_{t_n}^{\theta_n} h(X_s^{t_n,x_n})\,\mathrm{d}s + \psi(\theta_n, X_{\theta_n}^{t_n,x_n})\right].$$

Use Itô's formula between t_n and θ_n, divide by ϵ_n and use the bound in (A.13) to obtain

$$\frac{\gamma_n}{\epsilon_n} - \frac{\eta}{4} \leq \frac{1}{\epsilon_n}\mathbb{E}\left[\int_{t_n}^{\theta_n} (\partial_t\psi + \mathcal{L}\psi + h)(s, X_s^{t_n,x_n})\,\mathrm{d}s\right] \leq -\frac{\eta}{2}\mathbb{E}\left[\frac{\theta_n - t_n}{\epsilon_n}\right].$$

Let $n \to \infty$ to get $-\eta/4 \leq -\eta/2$, which is a contradiction. Thus the subsolution property follows. Finally, uniqueness is a direct consequence of the comparison result in Proposition A.8. □

Here, the lack of regularity of the value function precludes us from proving a verification-type result, as in the case of classical solutions to HJB equations, to prove that the strategy we find is indeed optimal. However, as shown by our results, the latency-optimal strategy outperforms the four benchmarks we consider (among which there is the robust and well-known TWAP), which lends strong support to our assumption that the random-latency strategy is the optimal strategy.

A.3 A note about numerical approximations

To compute a numerical approximation to the value function v_0, we proceed as follows. First, use the ansatz $v_0(t,x) = c + h_0(t,s,q)$ and compute an approximation to h_0 satisfying (A.10) with terminal condition $h_0(T,s,q) = (qs + q\zeta - aq(q-1))(1-\rho)$. Next, consider the grid discretisation $[0,T] \times [S_0 - 100\mathfrak{t}, S_0 + 100\mathfrak{t}] \times [0, Q_0]$ with 1000 points in the time axis, 201 points in the price axis and 11 in the inventory axis – here $\mathfrak{t} = 10^{-5}$ which is the tick-size in EUR/USD in LMAX. At the terminal time T, the value of the function is given by the terminal condition. Then for any point (t,s,q) in the grid, we approximate

$$\sup_{\ell\in L}\mathbb{E}\left[\Big((1-\rho)(S_{\tilde{\tau}-}^{t,s} + Z)f(\ell - S_{\tilde{\tau}-}^{t,s} - Z) + h_0\big(\tilde{\tau}, \tilde{\Gamma}(S_{\tilde{\tau}-}^{t,s}, Z, q, \ell)\big)\Big)\mathbf{1}_{\{\tilde{\tau}\leq T\}}\right] \quad \text{(A.18)}$$

together with the control $\ell^* \in L$ that attains the supremum in (A.18), which we obtain as follows. Split the set L into intervals of size 10^{-5} considering 100 ticks above and 100 ticks below $S_t = s$ – considering more ticks is not necessary because the probability that $\{|S_{\tilde{\tau}} - s| > 100 \times 10^{-5}\}$ is negligible. We perform (i) 10^4 simulations for $\tilde{\tau}$, where we use as $\hat{\tau} \in (t,T]$ the closest value to $\tilde{\tau}$ in the time grid, (ii) 10^4 simulations for $S_{\hat{\tau}}^{t,s}$ rounded to five decimals, and (iii) 10^4 simulations for Z. If for a given simulation, the fundamental best bid rate $S_{\hat{\tau}}^{t,s}$ is not in the grid, we use a cubic spline approximation with the functions $x \mapsto h(\hat{\tau}, x, q)$ and $x \mapsto h(\hat{\tau}, x, q-1)$. We store the optimal $\ell^* \in L$ as a function of (t,s,q) together with a variable that flags when an impulse happens.

The methodology described above is effective, but computationally expensive. However, once the value function is approximated, the implementation of the optimal strategy is fast enough so that it can be used to trade in the market. Further analysis of numerical methods for this type of problems is left for future research.

Appendix B: Benchmarks

B.1 Optimal execution with zero latency

Assume there is no latency in the marketplace. The trader controls the times at which she sends sell MOs to the exchange. The size of all MOs is one and the filled trades have permanent price impact as in the model of Sect. 3.

Let $\tau = (\tau_i)_{i\geq 1}$ denote an increasing sequence of stopping times when the trader sends sell MOs to the exchange. Observe that if latency is zero, a sell MO of size one at time τ_i is the same as a sell MLO for a unit size with a limit price $S^\tau_{\tau_i-}$ because the trader observes and acts with no delay in the marketplace and the exchange processes the order with no delay. The dynamics of the best bid price process, denoted by $S^\tau = (S^\tau_t)_{t\geq 0}$, follow

$$S^\tau_t = S_0 + \int_0^t b(S^\tau_u)\,\mathrm{d}u + \int_0^t \sigma(S^\tau_u)\,\mathrm{d}W_u - \kappa \sum_{\tau_i \leq t} 1,$$

where the functions $b, \sigma : \mathbb{R} \to \mathbb{R}$ are Lipschitz-continuous and κ is the permanent bid rate impact parameter. The cash process $C^\tau = (C^\tau_t)_{t\geq 0}$ and the inventory process $Q^\tau = (Q^\tau_t)_{t\geq 0}$ are given by

$$C^\tau_t = \sum_{\tau_i \leq t} S^\tau_{\tau_i-}, \qquad Q^\tau_t = \mathfrak{M} - \sum_{\tau_i \leq t} 1.$$

For each execution strategy τ over the trading window $[0, T]$, the trader computes

$$\Pi(\tau) = g(X^\tau_T) + \int_0^T h(X^\tau_s)\,\mathrm{d}s + \sum_{\tau_i \leq T} \mathfrak{C}(X^\tau_{\tau_i-}),$$

where $g, h : \mathbb{R}^3 \to \mathbb{R}$ are as in (3.9) and the function $\mathfrak{C}(s, q, c) = -\rho s$ represents a transaction cost every time a trade of unit size is executed. The trader solves the control problem $\sup_{\tau\in\mathcal{A}} \mathbb{E}[\Pi(\tau)]$, where

$$\mathcal{A} = \{\tau = (\tau_i)_{i\geq 1} : \text{for each } i \geq 1,\ \tau_i \text{ is an } \mathbb{F}\text{-stopping time, and } \tau_{i+1} > \tau_i\}.$$

For $(t, x) \in [0, T] \times \mathbb{R} \times (-\infty, \mathfrak{M}] \times \mathbb{R}$ and $\tau \in \mathcal{A}_t := \{\tau = (\tau_i)_{i\geq 1} \in \mathcal{A} : \tau_1 \geq t\}$, the system $X^{t,x,\tau}$ is given by

$$X^{t,x,\tau}_s = x + \int_t^s \boldsymbol{b}(X^{t,x,\tau}_u)\,\mathrm{d}u + \int_t^s \boldsymbol{\sigma}(X^{t,x,\tau}_u)\,\mathrm{d}W_u + \sum_{t<\tau_i\leq s} \big(\Gamma(X^{t,x,\tau}_{\tau_i-}) - X^{t,x,\tau}_{\tau_i-}\big),$$

where $\Gamma(s,q,c) = (s-\kappa, q-1, c+s)$. The trader's performance criterion and value function are

$$J(t,x,\tau) = \mathbb{E}\bigg[g(X_T^{t,x,\tau}) + \int_t^T h(X_s^{t,x,\tau})\,\mathrm{d}s + \sum_{\tau_i \le T} \mathfrak{C}(X_{\tau_i-})\bigg],$$

$$v(t,x) = \sup_{\tau\in\mathcal{A}_t} J(t,x,\tau).$$

Finally, for zero latency, the characterisations of the value function are standard results in the theory of impulse control, and we do not mention them here for brevity.

B.2 Optimal execution with deterministic latency

The trader employs MLOs and operates in the market with a fixed and known latency. The execution strategy has at most one pending order at any one time – we can easily extend this framework so the agent can have up to $m \in \mathbb{N}$ pending orders. Our approach is based on the work of Bruder and Pham [9], who develop a general framework for impulse control with deterministic delay.

Let $\Delta > 0$ be the fixed and known delay of the trader in the marketplace and $(Z_n)_{n\in\mathbb{N}}$ a collection of i.i.d. random variables with law as in (3.2) such that for $i \in \mathbb{N}$, Z_i is $\mathcal{F}_t$-measurable for $t \ge i\Delta$. As in Sect. 3, we use the sequence of marks $(Z_n)_{n\in\mathbb{N}}$ to model the flickers that affect the trader's MLO when the exchange processes the order.

The agent's execution strategy is $\beta = (\tau_i, \mathfrak{l}_i)_{i\ge1}$, where (τ_i) is an increasing sequence of stopping times at which the agent sends MLOs with limit price $\mathfrak{l}_i$. We require that $\tau_{i+1} - \tau_i \ge \Delta$ a.s. for all i, because the trader can have at most one pending order in the exchange. The set of admissible strategies is

$$\mathcal{A} = \{\beta = (\tau_i, \mathfrak{l}_i)_{i\ge1} : \text{for all } i \ge 1,\ \tau_{i+1} - \tau_i \ge \Delta \text{ a.s. and } \mathfrak{l}_i \text{ is } \mathcal{F}_{\tau_i}\text{-measurable}\}.$$

Interventions in the system $X_t^\beta = (S_t^\beta, Q_t^\beta, C_t^\beta)$ are at time τ_i, but processed by the exchange and notified to the trader at $\tau_i + \Delta$; so the system evolves from $X^\beta_{(\tau_i+\Delta)-}$ to $X^\beta_{\tau_i+\Delta} = \Gamma(X^\beta_{(\tau_i+\Delta)-}, Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor}, \mathfrak{l}_i)$, where Γ is as in (3.5) and $\lfloor\cdot\rfloor$ is the floor function. The value of $j = \lfloor(t+\Delta)/\Delta\rfloor$ is the index of the random variable Z_j which is measurable at time $j\Delta \in (t, t+\Delta]$.

Let $\beta = (\tau_i, \mathfrak{l}_i)_{i\ge1} \in \mathcal{A}$. The fundamental best bid price process is

$$S_t^\beta = S_0 + \int_0^t b(S_u^\beta)\,\mathrm{d}u + \int_0^t \sigma(S_u^\beta)\,\mathrm{d}W_u - \kappa \sum_{\tau_i+\Delta\le t} f(\mathfrak{l}_i - S^\beta_{(\tau_i+\Delta)-} - Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor}),$$

where f is as in the random latency case. The cash process satisfies

$$C_t^\beta = \sum_{\tau_i+\Delta\le t} f(\mathfrak{l}_i - S^\beta_{(\tau_i+\Delta)-} - Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor})(S^\beta_{(\tau_i+\Delta)-} + Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor}),$$

and the inventory is

$$Q_t^\beta = \mathfrak{M} - \sum_{\tau_i+\Delta\le t} f(\mathfrak{l}_i - S^\beta_{(\tau_i+\Delta)-} - Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor}).$$

For $X_0 = (S_0, \mathfrak{M}, 0)$, the controlled system X_t^β is the solution to the SDE

$$\begin{aligned} X_t^\beta &= X_0 + \int_0^t \boldsymbol{b}(X_u^\beta)\,\mathrm{d}u + \int_0^t \boldsymbol{\sigma}(X_u^\beta)\,\mathrm{d}W_u \\ &\quad + \sum_{\tau_i+\Delta\le t} \big(\Gamma(X^\beta_{(\tau_i+\Delta)-}, Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor}, \mathfrak{l}_i) - X^\beta_{(\tau_i+\Delta)-}\big), \end{aligned} \tag{B.1}$$

where the functions $\boldsymbol{b}, \boldsymbol{\sigma} : \mathbb{R}^3 \to \mathbb{R}^3$ are given by $\boldsymbol{b}(x_1, x_2, x_3) = (b(x_1), 0, 0)$ and $\boldsymbol{\sigma}(x_1, x_2, x_3) = (\sigma(x_1), 0, 0)$, and b, σ are Lipschitz-continuous. Finally, the function $\Gamma : \mathbb{R}^4 \times L \to \mathbb{R}^3$ is as in (3.5).

Now fix a finite horizon $T < \infty$ and assume that $T - \Delta \ge 0$ to avoid trivialities. As above, the agent computes

$$\Pi(\beta) = g(X_T^\beta) + \int_0^T h(X_s^\beta)\,\mathrm{d}s + \sum_{\tau_i+\Delta\le T} \mathfrak{C}(X^\beta_{(\tau_i+\Delta)-}, Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor}, \mathfrak{l}_i),$$

where $g, h : \mathbb{R}^3 \to \mathbb{R}$ and $\mathfrak{C} : \mathbb{R}^4 \times L \to \mathbb{R}$ are as in (3.9). Thus the agent solves the control problem $\sup_{\beta\in\mathcal{A}} \mathbb{E}[\Pi(\beta)]$, which is well posed if

$$\mathbb{E}\Big[\sup_{s\le T} |X_s^\beta|^2\Big] < \infty.$$

The system in (B.1) is non-Markovian. Given an impulse, the future of X^β is influenced by a possible pending order attempted between $t - \Delta$ and t. For any $t \in [0, T]$, $k = 0, 1$, we let $P_t(0) = \emptyset$ and

$$P_t(1) = \{p = (t_1, \ell_1) \in [0, T - \Delta] \times L : t - \Delta < t_1 \le t\}$$

denote whether there is a pending order in the exchange. For any $p = (t_1, \ell_1) \in P_t(1)$, $t \in [0, T]$, the set of admissible strategies at time t with pending order p is

$$\mathcal{A}_{t,p} = \{\beta = (\tau_i, \mathfrak{l}_i)_{i\ge1} \in \mathcal{A} : (\tau_1, \mathfrak{l}_1) = (t_1, \ell_1) \text{ and } \tau_2 \ge t_1 + \Delta\}.$$

Similarly, if there is no pending order in the exchange,

$$\mathcal{A}_{t,\emptyset} = \mathcal{A}_t = \{\beta = (\tau_i, \mathfrak{l}_i)_{i\ge1} \in \mathcal{A} : \tau_1 \ge t\}.$$

For any $(t, x) \in [0, T] \times \mathbb{R}^3$, $p \in P_t(k)$ for $k = 0, 1$ and $\beta \in \mathcal{A}_{t,p}$, we denote by $X^{t,x,p,\beta}$ the solution to

$$\begin{aligned} X_s^{t,x,p,\beta} &= x + \int_t^s \boldsymbol{b}(X_u^{t,x,p,\beta})\,\mathrm{d}u + \int_t^s \boldsymbol{\sigma}(X_u^{t,x,p,\beta})\,\mathrm{d}W_u \\ &\quad + \sum_{t<\tau_i+\Delta\le s} \big(\Gamma(X^{t,x,p,\beta}_{(\tau_i+\Delta)-}, Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor}, \mathfrak{l}_i) - X^{t,x,p,\beta}_{(\tau_i+\Delta)-}\big). \end{aligned}$$

Next, consider the performance criterion

$$J_k(t,x,p,\beta)=\mathbb{E}\Bigg[g(X_T^{t,x,p,\beta})+\int_t^T h(X_s^{t,x,p,\beta})\,\mathrm{d}s + \sum_{\tau_i+\Delta\le T}\mathfrak{C}(X^{t,x,p,\beta}_{(\tau_i+\Delta)-},Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor},\mathfrak{l}_i)\Bigg]$$

for $(t,x)\in[0,T]\times\mathbb{R}^3$, $p\in P_t(k)$, $k=0,1$, $\beta\in\mathcal{A}_{t,p}$, and the corresponding value functions

$$v_k(t,x,p)=\sup_{\beta\in\mathcal{A}_{t,p}} J_k(t,x,p,\beta),\qquad k=0,1 \text{ and } (t,x,p)\in\mathcal{D}_k,$$

where $\mathcal{D}_k=\{(t,x,p):(t,x)\in[0,T)\times\mathbb{R}^3, p\in P_t(k)\}$. Recall that for $k=0$, $P_t(0)=\emptyset$; therefore we write $v_0(t,x)=v_0(t,x,\emptyset)$, $\mathcal{D}_0=[0,T)\times\mathbb{R}^3$.

We state the DPP that holds for deterministic latency. Let $t\in[0,T]$, $\beta\in\mathcal{A}$, with $\beta=(\tau_i,\mathfrak{l}_i)_{i\ge1}$. Denote the number of executed orders by $\iota(t,\beta)$, the number of pending orders at time t by $k_t(\beta)$ (which is zero or one), and the pending order by $p(t,\beta)$ (which is the empty set if there is no pending order); so

$$\begin{aligned}\iota(t,\beta)&=\inf\{i\ge1:\tau_i>t-\Delta\}-1\in\mathbb{N}\cup\{\infty\},\\ k_t(\beta)&=\mathrm{card}\{i\ge1:t-\Delta<\tau_i\le t\}\in\{0,1\},\\ p(t,\beta)&=(\tau_{i+\iota(t,\beta)},\mathfrak{l}_{i+\iota(t,\beta)})_{1\le i\le k_t(\beta)}\in P_t\big(k_t(\beta)\big).\end{aligned}$$

For the proofs of the theorems that follow, see Bruder and Pham [9, Sects. 3 and 4].

Theorem B.1 *Let θ be any stopping time valued in $[t,T]$, possibly depending on β. For $k\in\{0,1\}$, $(t,x,p)\in\mathcal{D}_k$,*

$$v_k(t,x,p)=\sup_{\beta\in\mathcal{A}_{t,p}}\mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,p,\beta})\,\mathrm{d}s+\sum_{t<\tau_i+\Delta\le\theta}\mathfrak{C}(X^{t,x,p,\beta}_{(\tau_i+\Delta)-},Z_{\lfloor(\tau_i+\Delta)/\Delta\rfloor},\mathfrak{l}_i) + v_{k(\theta,\alpha)}\big(\theta,X_\theta^{t,x,p,\beta},p(\theta,\beta)\big)\Bigg].$$

The following corollary is a consequence of the DPP. For all $t\in[0,T]$, denote by $\mathcal{I}_t$ the set of pairs $(\tau,\mathfrak{l})$ where τ is a stopping time with $t\le\tau\le T-\Delta$ or $\tau=\infty$ a.s., and $\mathfrak{l}$ is an $\mathcal{F}_\tau$-measurable random variable in L.

Corollary B.2 *Let $(t,x)\in[0,T)\times\mathbb{R}^d$.*

(1) *For any stopping time θ valued in $[t,t_1+\Delta)$ and for $p=(t_1,\ell_1)\in P_t(1)$,*

$$v_1(t,x,p)=\mathbb{E}\Bigg[\int_t^\theta h(X_s^{t,x,0})\,\mathrm{d}s+v_1(\theta,X_\theta^{t,x,0},p)\Bigg].$$

(2) *For any stopping time* θ *valued in* $[t, t+\Delta)$,

$$v_0(t,x,\emptyset) = \sup_{(\tau,\mathfrak{l})\in\mathcal{I}_t} \mathbb{E}\Big[\int_t^\theta h(X_s^{t,x,0})\,\mathrm{d}s + v_0(\theta, X_\theta^{t,x,0}, \emptyset)\mathbf{1}_{\{\theta<\tau\}} + v_1\big(\theta, X_\theta^{t,x,0}, (\tau,\mathfrak{l})\big)\mathbf{1}_{\{\theta\geq\tau\}}\Big].$$

The proof is a straightforward modification of the proofs of Theorem 3.1 and Bruder and Pham [9, Corollary 3.2]. Define $\mathcal{D}_k^{1,2}$ for $k=0,1$ as $\mathcal{D}_0^1=(T-\Delta,T)\times\mathbb{R}^3$, $\mathcal{D}_0^2=[0,T-\Delta]\times\mathbb{R}^3$, $\mathcal{D}_1^1=\mathcal{D}_1$, $\mathcal{D}_1^2=\emptyset$.

Theorem B.3 *For* $k\in\{0,1\}$, $(t,x,p)\in\mathcal{D}_k^{1,2}$, *the value functions satisfy the dynamic programming equation*

$$\min\Big\{-\frac{\partial v_k}{\partial t}(t,x,p) - \mathcal{L}v_k(t,x,p) - h(x),\ v_k(t,x,p) - \sup_{\ell\in L} v_{k+1}\big(t,x,(t,\ell)\big)\Big\} = 0 \qquad \text{on } \mathcal{D}_0^2, \text{for } k=0, \tag{B.2}$$

and

$$-\frac{\partial v_k}{\partial t}(t,x,p) - \mathcal{L}v_k(t,x,p) - h(x) = 0 \qquad \text{on } \mathcal{D}_k^1, \text{for } k=0,1. \tag{B.3}$$

From the system of equations (B.2) and (B.3), we have the Feynman–Kac representation

$$v_1\big(t,x,(t_1,\ell_1)\big) = \mathbb{E}\Big[\int_t^{t_1+\Delta} h(X_s^{t,x})\,\mathrm{d}s + \mathfrak{C}(X_{(t_1+\Delta)-}^{t,x}, Z, \ell_1) + v_0\big(t_1+\Delta, \Gamma(X_{(t_1+\Delta)-}^{t,x}, Z, \ell_1)\big)\Big] \tag{B.4}$$

for all $(t_1,\ell_1)\in[0,T-\Delta]\times L$, $t\in[t_1,t_1+\Delta)\times\mathbb{R}^3$, and where Z is an independent random variable with law as in (3.2). Substitute (B.4) in (B.2) to obtain the variational inequality satisfied by v_0, and transform the original problem into a non-delay impulse control problem. Finally, we use the ansatz $v_0(t,x)=v_0(t,s,q,c)=c+h_0(t,s,q)$ to reduce the dimension of the value function by one.

Acknowledgements We are grateful to I. Cialenco, S. Cohen, R. Cont, R. Donnelly, L. Hughston, S. Jaimungal, J. Muhle-Karbe, J. Penalva, H. Pham, J. Ricci and A. Stewart for comments and discussions. We are grateful to seminar participants at SIAM SIAG/FME virtual seminars series, U. of Leeds and U. of Oxford.

Declarations

Competing Interests The authors declare no competing interests.

References

1. Alfonsi, A., Fruth, A., Schied, A.: Optimal execution strategies in limit order books with general shape functions. Quant. Finance **10**, 143–157 (2010)
2. Almgren, R.: Optimal execution with nonlinear impact functions and trading-enhanced risk. Appl. Math. Finance **10**, 1–18 (2003)
3. Almgren, R.: Execution costs. In: Cont, R. (ed.) Encyclopedia of Quantitative Finance, pp. 612–616. Wiley, New York (2010)
4. Almgren, R., Chriss, N.: Optimal execution of portfolio transactions. J. Risk **3**, 5–39 (2000)
5. Barger, W., Lorig, M.: Optimal liquidation under stochastic price impact. Int. J. Theor. Appl. Finance **22**, 1850059 (2019)
6. Bayraktar, E., Ludkovski, M.: Optimal trade execution in illiquid markets. Math. Finance **21**, 681–701 (2010)
7. Bertsekas, D.P., Shreve, S.: Stochastic Optimal Control: The Discrete-Time Case. Academic Press, San Diego (1978)
8. Bouchard, B., Touzi, N.: Weak dynamic programming principle for viscosity solutions. SIAM J. Control Optim. **49**, 948–962 (2011)
9. Bruder, B., Pham, H.: Impulse control problem on finite horizon with execution delay. Stoch. Process. Appl. **119**, 1436–1469 (2009)
10. Cartea, Á., Jaimungal, S.: Optimal execution with limit and market orders. Quant. Finance **15**, 1279–1291 (2015)
11. Cartea, Á., Jaimungal, S.: Incorporating order-flow into optimal execution. Math. Financ. Econ. **10**, 339–364 (2016)
12. Cartea, Á., Jaimungal, S., Penalva, J.: Algorithmic and High-Frequency Trading. Cambridge University Press, Cambridge (2015)
13. Cartea, Á., Jaimungal, S., Ricci, J.: Buy low, sell high: a high frequency trading perspective. SIAM J. Financ. Math. **5**, 415–444 (2014)
14. Cartea, Á., Arribas, I.P., Sánchez-Betancourt, L.: Double-execution strategies using path signatures. SIAM J. Financ. Math. **13**(4), 1379–1417 (2022). https://doi.org/10.1137/21M1456467
15. Cartea, Á., Sánchez-Betancourt, L.: The shadow price of latency: improving intraday fill ratios in foreign exchange markets. SIAM J. Financ. Math. **12**, 254–294 (2021)
16. Crandall, M.G., Ishii, H., Lions, P.L.: User's guide to viscosity solutions of second order partial differential equations. Bull. Am. Math. Soc. **27**, 1–67 (1992)
17. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Springer, Berlin (2006)
18. Gao, X., Wang, Y.: Optimal market making in the presence of latency. Quant. Finance **20**, 1495–1512 (2020)
19. Guéant, O.: Optimal execution and block trade pricing: a general framework. Appl. Math. Finance **22**, 336–365 (2015)
20. Guéant, O.: The Financial Mathematics of Market Liquidity: From Optimal Execution to Market Making. CRC Press, Boca Raton (2016)
21. Guéant, O., Lehalle, C.A., Fernandez-Tapia, J.: Optimal portfolio liquidation with limit orders. SIAM J. Financ. Math. **3**, 740–764 (2012)
22. Guilbaud, F., Pham, H.: Optimal high-frequency trading with limit and market orders. Quant. Finance **13**, 79–94 (2013)
23. Kalsi, J., Lyons, T., Perez Arribas, I.: Optimal execution with rough path signatures. SIAM J. Financ. Math. **11**, 470–493 (2020)

24. Moallemi, C.C., Saĝlam, M.: The cost of latency in high-frequency trading. Oper. Res. **61**, 1070–1086 (2013)
25. Øksendal, B., Sulem, A.: Optimal stochastic impulse control with delayed reaction. Appl. Math. Optim. **58**, 243–255 (2008)