Unraveling Limit Order Books Using Just Bid/Ask Prices

Jose Blanchet, Xinyun Chen and Yanan Pei

February 22, 2017

Abstract

How much of the structure of a Limit Order Book (LOB) by only observing the bid/ask price dynamics? In this paper we provide a model which, surprisingly, allows us to recover with reasonable empirical accuracy the general profile of the LOB in the setting of some small-tick stocks. Our approach exploits the empirically observed multi-scale dynamics of the LOB and we also apply such multi-scale analysis to obtain a jump diffusion model which connects the distribution of jumps with the distribution of orders inside the LOB.

1 Introduction

The Limit Order Book (LOB) is used in most of today’s financial markets to match buyers and sellers. In this light, LOB plays a substantial role in financial trading activities. The survey paper [10] provides a comprehensive review of the LOB, including its basic mechanisms, empirical features observed in different markets and LOB models developed in the literature.

As a trading venue, the LOB contains information about the potential sellers and buyers on the market, reflecting their perspectives on the value of the underlying security. Therefore, LOB information is important for the design of trading strategies (see, for example, [21]). However, the information of the whole LOB is not always available to the traders. In some cases, such as the equity market, the LOB data is not free; while in other cases, it is not open to the public at all. For example, in some electronic brokerage systems of the foreign exchange market, participants can only see the best price levels (the bid and ask prices) with associate order volume (see [20]). In contrast, trade and quote (TAQ) information, including the bid/ask price, sizes and prices of the trades, is much easier to access in financial markets.

The main contribution of this paper is to demonstrate that a substantial portion of the LOB persists at sufficiently large time scales. As a consequence of this persistence, and due to the high frequency of the trades that occur during such time scales, it is possible to recover a substantial amount of information of the LOB only from TAQ data. This is far from trivial, because the LOB contains much more information than the TAQ data. In particular, the LOB includes all available price levels for both the buy and sell sides, and the order volumes at each price level.

To perform this recovery, we start with a queueing model for the LOB dynamics. With the introduction of multiple time scales in the model, we are able to build a direct and analytic connection between the characteristics of the whole LOB and the bid/ask price process, observed at the time of trade. By such a connection, we are able to recover the characteristics of the LOB from the TAQ data which contains the bid/ask price information we need. Our approach is mostly useful for the so-called small-tick stocks: stocks which tend to have a persistent relatively large spread. As examples, we study 3 stocks, Amazon.com (ticker: AMZN), Goldman Sachs (ticker: GS) and Baidu, Inc (ticker: BIDU), to test our recovery method. In our numerical experiments, we show
that the average number of orders at each price level in the order book can be reconstructed with significant accuracy by observing basically the bid/ask price dynamics.

This insight already provides a substantial value to traders who do not have access to the entire LOB information. We believe that our ideas can be used in trading systems in which order books are maintained but only TAQ information is publicly available.

Moreover, we demonstrate in Section 6 how in long time scales (several minutes) the distribution of the LOB could manifest itself in the appearance of jumps in the limiting continuous-time price process. We derive a bivariate jump diffusion model for the joint evolution of the bid/ask prices, as observed at the times of trades in our LOB model. The volatility, jump distribution and intensity of the limiting process are linked to the structure of the LOB. To our best knowledge, this is the first paper that studies the connection between the price jumps and the order flow dynamics from a micro-structure / high-frequency perspective.

While our results are novel, the starting point of our approach, in particular a queueing model of the LOB dynamics, is closely related to various works in high frequency trading literature. For example, Markovian queue models are studied in [23], [6], [1], [5], [19], [18], [3] and [24]. Paper [2] generalizes the Markovian queues and uses Hawkes process to capture the autocorrelation in order flows as observed in real markets. Contributions [16] and [15] develop a class of state-dependent models to capture traders’ responses to the market depth. Given the complicated interplay between the order flow and price to change on the LOB, it is usually difficult to obtain an analytical expression of the price process statistics from the LOB queueing models. Hence, more restriction or model simplification are imposed to obtain analytic results on the price process. In [5], the authors consider only the bid and ask price levels of the LOB, and provide an expression of the price volatility in terms of the arrival and cancellation rates of orders. In [3] and [19], the authors consider only one side of the LOB to derive analytic characteristics of the price process. To recover the LOB structure from the TAQ information, however, we must connect the bid/ask price process with the whole LOB dynamics. To this end, we introduce three different time scales in our queueing model for the LOB:

(a) limit order event times (limit order arrivals and cancellations, fractions of a second),
(b) trade times (market order arrivals, usually around 10 seconds), and
(c) after many trades (a few minutes).

Our model is postulated at time scale (a), then we derive a simplified model at time scale (b), and finally we derive continuous-time model which is suitable for time scale (c). We emphasize that our time scalings are all motivated by empirical findings, as discussed in Section 2.2.

Under time scale (a), we model the dynamics of the LOB as a Markovian multi-class queueing system. The customers are the limit orders and they are waiting to be traded. This queueing system has two servers, corresponding to the bid and ask sides of the LOB. On each side, the service is provided by the market orders, as they cause transactions against limit orders immediately upon their arrivals, and the classes of the limit orders are determined by their prices. The highest service priority is given to the buy (sell) limit orders at the highest (lowest) price, and orders at the same price are served FIFO. While waiting, the limit orders can abandon the system at certain rates depending on their prices. These are the main ingredients of the queueing model. Of course, there are other aspects that are relevant and that we shall discuss in the body of the paper. For example, about the Markovian assumption (which implies Poisson arrivals), as we shall explain later, our analysis will allow us to ultimately accommodate self-exciting arrival processes and other extensions. At this moment, we only wish to convey the main elements to conceptually describe how our insights are obtained.

In the context of time scale (b) (the model at trade times), we observe empirically that the arrival rates of limit orders (i.e. orders inside the book) occur at a much faster rate than market
orders (i.e., trade transactions). We introduce a scaling that allows us to prove an averaging principle for the Markovian multi-class queueing system at the times of trades (i.e., at the arrival times of market orders). Such averaging principle allows us to compute (in the asymptotic limit) analytically the distribution of the bid/ask price changes between two consequent trades, as a function of the model parameters (such as the limit orders’ arrival rates and cancellation rates over the whole LOB). By these analytical results, we are able to provide a one-to-one correspondence between the bid/ask price changes, and the structure of the LOB.

Under time scale (c), we derive a jump-diffusion limit of the bid/ask price process in our LOB model. Interesting technical tools are needed to characterize the limiting process. For example, we have to introduce a local-time like “pushing” process in the description of the joint dynamics of the bid/ask price process to make sure that the spread is kept non-negative.

The paper is organized as follows. In Section 2, we discuss the queueing model of the LOB under time scale (a). In Section 3, we do an asymptotic analysis of the queueing model under time scale (b), and obtain the analytic results connecting the price change distribution and the LOB structure. In Section 4, we use the analytic results to study the daily trading data of AMZN, GS and BIDU. In Section 5, we discuss various extensions which can be easily accommodated within our multiscale analysis framework. Finally, in Section 6, we derive the macroscopic price dynamics on the LOB characterized by a pair of jump-diffusion processes under time scale (c).

2 The High Frequency Model

2.1 Model and Assumptions

In our queueing model, the LOB at time $t$ is represented by vector $q(t)$ such that the $i$-th coordinate for $i > 0$, $q_i(t)$, is the number of sell limit orders that are waiting in the LOB at time $t$ at price $i\delta$. The number of buy limit orders at $i\delta$ are represented with a negative sign $q_i(t)$. We shall impose assumptions to make sure that we do not have a situation in which $q_j(t) < 0$ and $q_i(t) > 0$ for $i < j$ because this will mean that there are standing orders to be sold at a price, $i\delta$, which is lower than the price $j\delta$ at which $|q_j(t)|$ buy orders are standing in the book.

The parameter $\delta$ is the so-called tick size of the LOB. For most U.S. listed stocks, $\delta$ equals to 1 cent. Figure ?? illustrates the representation of $q(t)$.

For each limit order (ask or bid, respectively) we define its relative price as the difference between its own price and the corresponding benchmark (ask or bid) prices, as we shall specify when we describe the arrivals and cancellations of limit orders.

The benchmark price of all sell limit orders at time $t$ are computed as follows. First, define

$$a(t) := a(q(t)) := \min\{i\delta > 0 : q_i(t) > 0\},$$

if $q_i(t) \leq 0$ for all $i > 0$, then $a(q(t)) = \infty$. Next define

$$b(t) := b(q(t)) := \max\{i\delta > 0 : q_i(t) < 0\},$$

if $q_i(t) \geq 0$ for all $i > 0$ then set $b(q(t)) = -\infty$.

In our model, the dynamics of the limit orders will depend on their relative prices, instead of their actual prices. As a limit order’s relative price changes with the benchmark prices, this dependence reflects traders’ response towards the change of market condition.

We shall assume that all types of orders are of the same size, or equivalently, of one unit of shares. The dynamics of a LOB is driven by the following three types of events:
Figure 1: Illustration of the Queueing Model for LOB at time $t$

1. Limit order arrivals: a new limit order is inserted to the LOB.

2. Limit order cancellations: an existing limit order is removed from the LOB.

3. Market order arrivals: a market order gets immediately executed with an opposite limit order sitting at the best price level upon its arrival.

Now we shall describe the stochastic processes used in our model corresponding to the three types of events.

**Arrivals of Market Orders:**

In order to capture the autocorrelation observed in the time series of trading volumes on both sides, we shall model the arrivals of market orders by a 2-dimensional Hawkes processes. In particular, let $\mu^a(t)$ ($\mu^b(t)$) be the arrival rate of sell (buy) market orders and $N^a(t)$ ($N^b(t)$) be the total number of sell (buy) market orders that have arrived by time $t$. The arrival rates at $t \geq 0$ satisfy

$$\mu^a(t) = e^{-\kappa t} \left( \mu^a(0) + \mu \int_0^t e^{\kappa s} s + \int_0^t \delta_1 e^{\kappa s} dN^a(s) + \int_0^t \delta_2 e^{\kappa s} dN^b(s) \right),$$

$$\mu^b(t) = e^{-\kappa t} \left( \mu^b(0) + \mu \int_0^t e^{\kappa s} s + \int_0^t \delta_2 e^{\kappa s} dN^a(s) + \int_0^t \delta_1 e^{\kappa s} dN^b(s) \right),$$

(1)
with $\kappa > 0$, $\mu > 0$, $\delta_1, \delta_2 > 0$. To avoid explosive behavior in the market order arrival rates, we assume $\delta_1 + \delta_2 < \kappa$, see [13] for stability properties of (1).

Let $\{t_k\}$ be the sequence of market order arrival times. The dynamics of the limit orders is described as follows:

**Arrivals and Cancellations of Limit Orders:**

L1. Arrivals of limit sell and buy orders are modeled as two independent Poisson processes with rates $\lambda^a$ and $\lambda^b$, respectively.

L2. During the time interval $[t_k, t_{k+1})$, the benchmark prices for sell and buy limit orders are $a(t_k)$ and $b(t_k)$ respectively. Note that a sell limit order placed at relative price of $i\delta$ at time $t$ means that its price equals to $a(t_k) + i\delta$, and a buy limit order placed at relative price of $i\delta$ at time $t$ means that its price equals to $b(t_k) - i\delta$.

L3. A limit sell (buy) order, arriving during the time interval $[t_k, t_{k+1})$, will choose its relative price equal to $i\delta$ with probability $p^a(i\delta; a(t_k), b(t_k)) \cdot (p^b(i\delta; a(t_k), b(t_k)))$ upon its arrival.

L4. We assume that $p^a(i\delta; a, b) = p^b(i\delta; a, b) = 0$ for all $i\delta \leq -(a - b)/2$, so that the actual price of an arriving limit sell (buy) order, namely $a + i\delta (b - i\delta)$ is greater (less) than $(a + b)/2$. Consequently, we have that

$$\sum_{i\delta > -(a-b)/2} p^a(i\delta; a, b) = 1 = \sum_{i\delta > -(a-b)/2} p^b(i\delta; a, b).$$

L5. During the time interval $[t_k, t_{k+1})$, sell (buy) limit orders waiting at relative price of $i\delta$ are cancelled at rate $\alpha^a(i\delta; a(t_k), b(t_k)) > 0$ ($\alpha^b(i\delta; a(t_k), b(t_k)) > 0$).

L6. The arrivals of limit orders and the price selection mechanism of the limit orders are mutually independent. Limit order arrivals are also independent of market order arrivals.

**Remark:** Our treatment allows to weaken the assumption that all market orders are of size 1 and allow market orders which arrive in batches, as long as the size of the incoming market order batch is less than the volume of standing limit orders at the best price level. Additional extensions, which accommodate more complex dependence between the arrivals of market and limit orders is studied in Section 5.

Assumption L2 and L4 play important roles in the model analysis. Under Assumption L4, the ask and bid sides of the LOB can be treated as two separate systems. Under Assumption L2, between the arrivals of two consequent market orders, the benchmark prices remain unchanged. As a result, each side of the LOB evolves just as a set of independent infinite-server queues. Each infinite-server queue corresponds to a price level, and the queue consists of the limit orders at that price level. The arrival rate of the infinite-server queue is just the arrival rates of limit orders at the corresponding price level, and the “service rate” is equal to the cancellation rate of the corresponding class.

Before we move on into the model analysis part, we shall briefly discuss the empirical support for us to impose these assumptions.
2.2 Discussion of model assumptions

Assumption L2 states that the arrival and cancellation rates of limit orders depend on the bid/ask prices at the time of the previous trade, instead of depending on the real-time bid/ask price, as in the models studied in [6], [16] and [2]. We impose this assumption in order to model a scenario in which market participants retain a rich amount of market information (i.e. the actual prices during transactions) while trying to reduce variability on the amount of information used to post and cancel orders. The real-time bid/ask price of limit orders can change without trade and tend to be noisy. Indeed, due to the growing popularity of algorithmic trading, limit orders are put and cancelled without being executed at high frequency, especially for those inside the spread ([12]). Therefore, the continuous observation of the bid/ask prices may result in a process with too much microstructure “noise” due to variability caused by cancellations of such fleeting orders. In the rest of the paper, we shall call the bid/ask price process observed at the times of trades the bid/ask price per trade.

Assumption L4 states that the arriving limit orders never cross the mid-price at the time of the previous trade. The assumption largely simplifies our model analysis. We believe that this assumption is reasonable for the so-called small-tick stocks we shall study. This is because, when the spread is persistently large, it is expensive to put limit orders crossing the mid-price. In our data, we find that the proportion of limit orders which cross the mid-price is small for the three stocks we study. Table 1 gives the average percentage of cross mid-price limit orders for AMZN, BIDU and GS among the 20 trading days in May of 2011. GS has a tighter spread and hence its cross-mid-price percentage is slightly larger.

Table 1: Descriptive Statistics

<table>
<thead>
<tr>
<th>Statistics (Average of May 2011)</th>
<th>AMZN</th>
<th>BIDU</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Limit Order Submission</td>
<td>311143.10</td>
<td>299579.57</td>
<td>217414.62</td>
</tr>
<tr>
<td>Number of Trades</td>
<td>22386.71</td>
<td>23661.71</td>
<td>13319.29</td>
</tr>
<tr>
<td>Percentage of Hidden Trades</td>
<td>20.53%</td>
<td>22.08%</td>
<td>16.62%</td>
</tr>
<tr>
<td>Tick Size (cents)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Average Spread (cents)</td>
<td>7.41</td>
<td>6.11</td>
<td>3.83</td>
</tr>
<tr>
<td>Daily Cross Mid-Price Limit Order Percentage</td>
<td>4.93%</td>
<td>5.31%</td>
<td>7.69%</td>
</tr>
<tr>
<td>Cancel Without Execution LO Percentage</td>
<td>95.50%</td>
<td>94.48%</td>
<td>95.39%</td>
</tr>
</tbody>
</table>

3 Averaging Principle: Connecting Order Flows with Price Returns

In this section, we derive an analytic characterization of the bid/ask price per trade processes in terms of the model parameters. The result is obtained using the stochastic-average principle under an asymptotic regime as suggested by empirical data. We shall first report the empirical observations that motivate our asymptotic analysis.

3.1 Empirical observations

Empirical Observation 1: Multi-Scale Evolution of Limit Order Flows and the Occurrence of Trades.
The contrast between the daily number of limit order events (including the submission and cancellation of limit orders) and the daily number of trades has been documented in LOB literature (see, for instance, [7]). We also summarize our empirical observation on the AMZN, BIDU and GS data in Table 1 and provide basic descriptive statistics which support the fact that the evolution of limit orders is much more frequent than the arrivals of market orders in the LOB.

**Empirical Observation 2: Fleeting Orders and High Cancellation Rate.**

Due to the prevalence of algorithmic trading in these days, the cancellation rate of limit orders has increased dramatically over recent years. For example, [11] reports the cancellation rate has increased dramatically, while [12] finds that in the year of 2010, about 95% of the limit orders were canceled without execution in the NASDAQ market. Table 1 reports the monthly averages of percentages of limit orders that are cancelled without (partial) execution in AMZN, BIDU and GS during May of 2011, and it also illustrates the high cancellation rate phenomenon in the most recent high frequency trading environment.

The high cancellation rate can be attributed to the large proportion of “fleeting” limit orders, which are usually put inside the spread and then canceled immediately if not being executed right away. Hautsch and Huang [12] reports that in 2010 NASDAQ data, the mean inter-arrival time of market orders is about 7 seconds while the mean cancellation time of limit orders inside the spread is less than 0.2 second.

### 3.2 Price change distribution

Before stating the formal results on the price change distribution, we briefly discuss the intuition behind the results and the proofs. Under Assumption L4, we can analyze the bid and ask sides of the LOB separately. So we shall focus on the ask side in the discussion, since the bid side follows the same argument.

Following the empirical observations, we will assume that the arrival rates $\lambda^a$ and cancellation rates $\alpha^a(\cdot)$ are substantially higher than the arrival rates of market orders $\mu^a(\cdot)$ in our model. We now want to study the distribution of the ask price changes per trade, i.e. $a(t_{k+1}) - a(t_k)$, where $\{t_k\}$ are the time of trades. Under our assumptions, between the arrivals of two consecutive market orders, the dynamic of the ask side of the LOB is equivalent to a set of independent infinite-server queues. With $\lambda^a >> \mu^a(\cdot)$ and $\alpha^a(\cdot) >> \mu^a(\cdot)$ as suggested by the Empirical Observations 1 and 2, heuristically, we can approximate the distribution of the queues in the LOB at time $t_{k+1}$ (given the state of the system at time $t_k$) by the associated steady-state distribution.

Note that between time $t_k$ and $t_{k+1}$, the infinite-server queue of limit orders at relative price $i\delta$ has arrival rate $\lambda^a(\cdot; a(t_k), b(t_k))$ and cancellation rate $\alpha^a(i\delta; a(t_k), b(t_k))$. As a result, the steady-state distribution of the number of standing limit orders at relative price $i\delta$ is Poisson, with the mean equal to the arrival rate $\frac{\lambda^a(i\delta; a(t_k), b(t_k))}{\alpha^a(i\delta; a(t_k), b(t_k))}$. Using this heuristic approximation, we have that

$$P(a(t_{k+1}) - a(t_k) \geq i\delta) = P(\text{The queues at relative prices lower than } i\delta \text{ are all empty at time } t_{k+1})$$

$$\approx \exp \left( - \sum_{j\delta \geq -(a(t_k) - b(t_k))/2}^{i-1} \frac{\lambda^a(j\delta; a(t_k), b(t_k))}{\alpha^a(j\delta; a(t_k), b(t_k))} \right) \triangleq \theta^a(i\delta; a(t_k), b(t_k)).$$

(2)
Following the same argument, we can also approximate

\[
P(b(t_{k+1}) - b(t_k) \leq -i\delta) \approx \exp \left( - \sum_{j\delta \geq -(a(t_k) - b(t_k))/2} \frac{\lambda^b(j\delta; a(t_k), b(t_k))}{\alpha^b(j\delta; a(t_k), b(t_k))} \right) \triangleq \theta^b(i\delta; a(t_k), b(t_k)).
\]

(3)

Since we are mainly interested in the price changes at the time of trades, we use \(\bar{a}(t)\) (\(\bar{b}(t)\)) to denote the ask (bid) price per-trade, i.e., \(\bar{a}(t)\) (\(\bar{b}(t)\)) equals to the ask (bid) price at the time of the last trade by time \(t\). The approximations (2) and (3) are equivalent to saying that the processes \(\{\bar{a}(t) : t \geq 0\}\) and \(\{\bar{b}(t) : t \geq 0\}\) jump every time when a trade occurs and the jump sizes almost follow the distributions given by \(\theta^a\) and \(\theta^b\). This heuristic can be validated by the stochastic-averaging principle ([17]) and the result is summarized in Theorem 1. We use \(D(0, \infty)\) to denote the space of right-continuous with left limits functions from \([0, \infty)\) to \(\mathbb{R}\) endowed with the Skorokhod \(J_1\) topology (see [4] for reference).

**Theorem 1.** Consider a sequence of LOB systems indexed by \(n\). In the \(n\)-th system, the total number of orders in the order book is given by \(q_n(0) = q(0) < \infty\) at time \(0\), and we let \(a_n(0) = a(0) = a\) and \(b_n(0) = b(0) = b\). We assume that the market orders arrive according to the Hawkes process (1) and the distribution of incoming limit orders is \(p^a_n(\cdot) = p^a(\cdot)\) and \(p^b_n(\cdot) = p^b(\cdot)\) (i.e. the distribution remains constant along the sequence of systems). Suppose there exists a sequence of positive numbers \(\{\xi_n : n \geq 1\}\) such that \(\lambda^a_n = \xi_n\lambda^a, \lambda^b_n = \xi_n\lambda^b, \alpha^a_n(\cdot) = \xi_n\alpha^a(\cdot), \alpha^b_n(\cdot) = \xi_n\alpha^b(\cdot)\) and \(\xi_n \to \infty\) as \(n \to \infty\).

We also assume the regularity condition that for any \((a, b) \in \mathbb{Z}^2\)

\[
\lim_{i \to \infty} \theta^a(i\delta; a, b) = 0, \text{ and } \lim_{i \to \infty} \theta^b(i\delta; a, b) = 0.
\]

(4)

Then, the corresponding price per-trade process \((\bar{a}_n(\cdot), \bar{b}_n(\cdot))\) converges weakly in \(D([0, \infty), \mathbb{Z}^2)\) to a pure jump process \((\bar{a}(\cdot), \bar{b}(\cdot))\). The process \((\bar{a}(\cdot), \bar{b}(\cdot))\) jumps at times corresponding to the arrivals in the Hawkes process (1). Moreover, if \(t\) is an arrival time in the Hawkes process, then

\[
P(\bar{a}(t) - \bar{a}(t-) = i\delta, \bar{b}(t) - \bar{b}(t-) = -j\delta | \bar{a}(t-), \bar{b}(t-))
\]

\[
= [\theta^a(i\delta; \bar{a}(t-), \bar{b}(t-)) - \theta^a((i + 1)\delta; \bar{a}(t-), \bar{b}(t-))]
\]

\[
\times [\theta^b(j\delta; \bar{a}(t-), \bar{b}(t-)) - \theta^b((j + 1)\delta; \bar{a}(t-), \bar{b}(t-))].
\]

**Remark:** The regularity condition (4) on \(\theta^a(i\delta; a, b)\) not only is quite natural, but it also can be easily verified in terms of \(p^a(i\delta; a, b)\) and \(\alpha^a(i\delta; a, b)\) by the explicit formula for \(\theta^a(i\delta; a, b)\) given in equation (2). The regularity condition on \(\theta^b(i\delta; a, b)\) holds following (3).

It is important to note that in the previous result we do not scale the arrival processes of market orders, so this result simply describes the price processes at time scales corresponding to the inter-arrival times of market orders (i.e. in the order of a few seconds according to the representative date discussed earlier). In Section 6 we shall introduce a scaling that will allow us to consider the process in longer time scales (several minutes or longer) by increasing the arrival rate of market orders.
4 Connecting Order Flows with Price Changes: An Empirical Study

In this section, we use our model and Theorem 1 to study the limit order book data of AMZN, BIDU and GS. AMZN and BIDU are listed on NASDAQ, and GS is listed on NYSE. They are all small-tick stocks with daily average spreads being 7.41 cents, 6.11 cents and 3.83 cents in May 2011, respectively (as reported in Table 1).

4.1 The LOB data

We use AMZN, BIDU and GS limit order book data up to level 50 for 20 consecutive trading days during May 2011 from NASDAQ’s Historical TotalView-ITCH data. For each stock, the data contains information of all LOB events, including the submissions, cancellations and executions of the limit orders at the first 50 price levels during the specified time period. All the LOB events are time stamped from 9:30 am to 16:00 pm in EST, with decimal precision of nanoseconds ($10^{-9}$ second).

For each trading day we remove the data in the first half hour after market opens (9:30 am - 10:00 am in EST) and the last half hour before market closes (15:30 pm - 16:00 pm in EST), as the LOB dynamics is more volatile in these two period than in the rest of the trading day. We also compress consecutive trades that have the same order ID, same price, at the same direction (bid or ask) and with time difference less than 1 millisecond as a single trade, in order to reduce the noise in model calibration.

4.2 Model validation: empirical test of Theorem 1

In this section we use the LOB data to test Equations (2) and (3) derived from our model. We first observe the empirical tail probability of the price change per-trade from the data. Then, we calculate the model-implied tail probability of price change per-trade by plugging in Equations (2) and (3) the calibrated arrival and cancellation rates of limit orders at all relative price levels. The detailed calibration steps are given in Appendix 8.3.

Figure 2 compares the daily empirical and model-implied tail probabilities of the price change per-trade on the bid and the ask sides, respectively, during 20 trading days in May of 2011. Figure 3 reports the same results for BIDU, and Figure 4 for GS, in the same time window.

We can see that the empirical and model predicted tail probabilities agree quite well at positive relative price levels. However for relative price level $i\delta \leq 0$, our model tends to give lower estimates compare to the empirical results, which might be due to the lack of hidden limit order information in the data base. A detailed discussion on our treatment of hidden orders is given in Appendix Section 8.3.

4.3 Model application: recover LOB from the price change distribution

We now apply our model to estimate the expected volumes of visible limit orders at each relative price level, upon the arrival of an market order, using only the bid/ask price per trade data.

According to Theorem 1, at each price level $i\delta$, either on the bid or ask side, the number of limit orders can be approximated by the steady-state distribution of the infinite server queue, which is a Poisson random variable with mean $\rho^{a/b}(i\delta; a, b) := \lambda^{a/b} \rho^{a/b}(i\delta; a, b) / \alpha^{a/b}(i\delta; a, b)$. On the other hand, by Equations (2) and (3), we have

$$\rho^{a/b}(i\delta; a, b) = \log \left( \theta^{a/b}(i\delta; a, b) \right) - \log \left( \theta^{a/b}((i + 1)\delta; a, b) \right).$$

(5)
Figure 2: AMZN tail probabilities in May 2011
Figure 3: BIDU tail probabilities in May 2011
Figure 4: GS tail probabilities in May 2011
Figure 5: AMZN LOB volume estimation results for both sides in May 2011
Figure 6: BIDU LOB volume estimation results for both sides in May 2011
Figure 7: GS LOB volume estimation results for both sides in May 2011
Note that $\theta^{a/b}(\cdot)$ is the tail probability of the bid/ask price change per-trade and can be directly observed in the bid/ask price per trade data. Therefore, we first obtain the empirical estimation for $\theta^{a/b}(i\delta; a, b)$ for each relative price $i\delta$ from the bid/ask price per trade data, and then compute the estimation for the mean number of limit orders sitting at $i\delta$ using equation (5). Since we assume that all orders are of the same size in our model, we shall estimate the mean volume as the mean number of orders multiplied by the average size of market orders, or equivalently, the average trade size.

For each trading day, we do the estimation for the relative price levels that are smaller than the 95% quantile of the empirical price change per-trade distribution. The estimation results for AMZN, BIDU and GS are reported in Figure 5, 6 and 7, with a comparison to the actual mean volumes observed in the LOB upon the arrivals of the market orders. Through the graphs, we can see that our estimation of the limit order volume is good, especially for the relative price levels around 0.

To statistically test how good our model performs in terms of recovering the visible limit order volumes at trade times, we run linear regressions

$$V^{bid/ask}_{model} = \beta^{bid/ask}_0 + \beta^{bid/ask}_1 \cdot V^{bid/ask}_{emp} + \epsilon^{bid/ask},$$

where $V^{bid/ask}_{model}$ is the vector of model estimated average volumes at relative price levels $-2, -1, 0, 1, 2$ during 20 trading days in May 2011, and $V^{bid/ask}_{emp}$ is the vector of empirical average volumes at the same relative price levels in the same time range.

For stock AMZN, we can see from Figure 8 that for both bid and ask sides, the linear regressions have high $R^2$ (both above 85%) and with 95% confidence level we cannot reject null hypothesis $\beta^{bid/ask}_0 = 0$ and $\beta^{bid/ask}_1 = 1$. Similar test results for stocks BIDU and GS are given in Figure 9 and Figure 10. Although statistically we will reject the null hypothesis $\beta^{bid/ask}_1 = 1$ for these cases, yet the upper bounds of the 95% confidence intervals of $\beta^{bid/ask}_1$ are quite close to 1.

Also note that our estimated average volumes are highly correlated (correlation coefficients above 90%) with the empirical average volumes for all three stocks, in that sense our model is able to recover the volumes of limit orders sitting at different price levels to a substantial extent, even though we may fail the $t$ tests in the previous linear regression setting in some cases.

The empirical results indicate that our model is able to recover the limit order volumes on several relative prices by observing the TAQ data only. It might be surprising that the recovery is remarkably good despite the simple independence assumption of limit orders at different price levels, which we have imposed to make the model tractable.

To check if our model could adapt to different situations as stock trading dynamics change, we also perform a similar study for AMZN in a 20-day time window from 2014-05-01 to 2014-05-29. Compared with May 2011, the stock price of AMZN has increased from around $200 to around $300, and the trading activity becomes less active. For the time period we choose in 2014, the average spread is 19.60 cents, much wider than what we have in May 2011, 7.41 cents. Figure 11 shows the empirical and model estimated average volumes of visible limit orders at each relative price level upon the arrival of a market order. Again our model gives satisfactory results despite the change of trading environment.

### 4.4 Discussion

In terms of recovering the information inside the LOB just by observing the bid/ask price, our approach seems to perform reasonably well. Still, deviations arise from the actual order book and the model prediction during a few days in the testing trading month. It is appropriate to close this
Figure 8: AMZN Volume Estimate Fitness Check in May 2011

Figure 9: BIDU Volume Estimate Fitness Check in May 2011
section with a discussion on the limitations and the possible issues in our approach that might play a role in some of the salient discrepancies.

The first possible issue that comes to our mind is the condition for the averaging principle to hold. In other words, it might be the case that the dynamic of the limit order events does not reach stationarity over the inter-trade intervals and hence the steady-state approximation is too crude. We attempted to test this potential problem by splitting each and everyone of the inter-trade intervals into five subintervals of equal size each. In Theorem 1, the analytical expression of the price change distribution mainly depends on the ratio of the arrival rates and cancellation rates of limit orders. Therefore, in Figure 12, we report the ratios between the frequency of limit order arrivals and cancellations observed in each of the five sub-intervals during May 2011 for different price levels. If the ratios remain constant across the five subintervals then the data is consistent with the application of the averaging principle.

In general, we see that, for relative price levels \( \geq 0 \), the limit order events remain constant over the five subintervals within the inter-trade intervals. For relative price levels \( < 0 \), the limit order events are less uniform, but the difference is moderate. So, this issue by itself did not appear to significantly explain discrepancies.

Given that most of the significant discrepancies happen on the relative price levels \( < 0 \), it is quite possible that the existence of hidden orders might have significant impact in our analysis. This problem is specially challenging since there is no direct information on the arrivals and cancellations of hidden orders in our data. We believe that we have dealt with this challenge within the constrains of our model in a reasonable way (see Appendix 8.3), but certainly the fact that hidden orders cannot be fully observed was, we believe, a major contributing factor in the observed discrepancies.

Another issue that we do not model is that the system is assumed to be time homogeneous throughout the day. We only remove the first and last half hours of the trading day to mitigate non-homogeneous components. But relevant pieces of news might have an effect which changes the dynamics significantly and this will not be captured by our model. To solve this issue requires more sophisticated statistical analysis on the LOB data jointly with news reports data, which is out of the scope of the current paper.
Figure 11: AMZN LOB volume estimation results for both sides in May 2014
Figure 12: Inter-trade stationarity check for AMZN, BIDU and GS on both sides in May 2011
5 Model Extension

The empirical results in Section 4 show that our multi-scale model yields reasonably accurate recovery despite the independence assumptions between the arrival of limit and market orders. In this section, we show how the multi-scale framework can be extended in order to accommodate a dependence characteristics which may reflect the interaction between market participants in financial markets.

For instance, a trader who wants to buy or sell the security has a choice between placing a market order or a limit order. This choice, which may be a function of the whole order book, can induce a general dependence structure. In the following extended model, we allow the traders to choose between market and limit orders. Moreover, we allow traders’ choices of the order types (market or limit) and of the limit order prices all to be dependent on the current LOB status $q(t)$. Recall that $q(t) = \{q_i(t)\}$ has been defined in Section 2.1 with $q_i(t)$ representing the number of limit orders at price $p(t) + i\delta$ in the LOB. Thus, the buy and sell sides of the LOB can be correlated as the order flow dynamics on both sides depends on the same vector $q(t)$.

Given that the dynamic of the LOB now is much more complicated than the model that we have studied in Section 3, we shall impose an upper bound on the prices at which limit order can be placed in the order book, in order to guarantee the regularity conditions as required in Kurtz [17] when applying the stochastic averaging principle. Although the assumption is imposed for technical reasons, we believe that this assumption can be motivated from a practical point of view. In practice, there are protocols in place, such as the “limit up-limit down” mechanism approved by SEC on US equity market (SEC [22]), in the operation of LOB dynamics which prevent extremely wide price fluctuations during a single trading day. So this sort of mechanism motivates our technical assumption, practically speaking.

To be precise, the LOB status is $\{q_i(t)\}_{i=1}^I$, where $\Gamma\delta$ is the maximum price at which limit orders can be placed. Now, given $q(t)$, the ask price of the LOB is defined as

$$a(t) = a(q(t)) = \min(\min\{0 < i\delta \leq \Gamma\delta : q_i(t) > 0\}, \Gamma + 1),$$

and the bid price as

$$b(t) = b(q(t)) = \max(\max\{i\delta > 0 : q_i(t) < 0\}, 0).$$

The mathematical description of the extended model is given as follows:

**Extended model: dynamics of the limit and market orders**

1. Arrivals of sell and buy orders are modelled as a independent Poisson process with rates $\lambda^a > 0$ and $\lambda^b > 0$.

2. For a sell (or buy) order arriving at time $t$, suppose the last trade happens at time $t_k < t$.

   Then, with probability $p^{m,a}(a(t_k), b(t_k), q(t))$ (or $p^{m,b}(a(t_k), b(t_k), q(t))$), the order is chosen by the trader to be a market order order and with probability $1 - p^{m,a}(a(t_k), b(t_k), q(t))$ (or $1 - p^{m,b}(a(t_k), b(t_k), q(t))$) to be a limit order.

3. If the sell (or buy) order is chosen to be a limit order. Then, it will choose its relative price, with respect to $a(t_k)$ (or $b(t_k)$) equal to $i\delta$ with probability $p^a(i\delta; a(t_k), b(t_k), q(t))$ (or $p^b(i\delta; a(t_k), b(t_k), q(t))$).
4. We assume that \( p^\alpha(i\delta; a(t_k), b(t_k), q(t)) = 0 \) for all \( i\delta \leq b(t_k) - a(t_k) \) and \( p^\beta(i\delta; a(t_k), b(t_k), q(t)) = 0 \) for all \( i\delta \leq b(t_k) - a(t) \), so that there is no trade triggered by an incoming limit order. We also assume that if \( i > \Gamma \)

\[
p^\alpha(i\delta; a(t_k), b(t_k), q(t)) = 0 = p^\beta(i\delta; a(t_k), b(t_k), q(t)).
\]

5. During the time interval \([t_k, t_{k+1})\), limit sell (or buy) orders waiting at relative price of \( i\delta \) are cancelled at rate \( \alpha^\alpha(i\delta; a(t_k), b(t_k), q(t)) > 0 \) (or \( \alpha^\beta(i\delta; a(t_k), b(t_k), q(t)) > 0 \)). We also assume that if \( i > \Gamma \)

\[
\alpha^\alpha(i\delta; a(t_k), b(t_k), q(t)) = 0 = \alpha^\beta(i\delta; a(t_k), b(t_k), q(t)).
\]

6. The orders choose their types and limit prices independently.

7. For any fixed \((\bar{a}, \bar{b})\), consider a modified process (abusing notation slightly), \( q(t) \), with constant arrival rates \( \lambda^{\alpha/b} p^\alpha(i\delta; \bar{a}, \bar{b}, q(t)) \) and cancellation rates \( \alpha^{\alpha/b}(i\delta; \bar{a}, \bar{b}, q(t)) \). We assume that the modified process \( q(\cdot) \), for fixed \((\bar{a}, \bar{b})\), is an irreducible and positive recurrent continuous time Markov chain.

Remark: As a consequence of Assumption 7, for each fixed \((a(t_k), b(t_k))\) the (modified) LOB process \( q(\cdot) \) with no market order arrival after time \( t_k \) possesses a unique stationary distribution \( \pi(dq; a(t_k), b(t_k)) \).

In the asymptotic regime, we will send the arrival rates of all types of orders to infinity, but the probability that the traders choose market orders to zero. This is consistent with the empirical observation that trade happens much less frequently than limit order events. Under this asymptotic regime the stochastic average principle holds and we have the following convergence result for the extended model.

**Theorem 2.** Consider a sequence of LOBs indexed by \( n \). In the \( n \)-th system, \( \lambda^a_n = \xi_n \lambda_a \), \( \lambda^b_n = \xi_n \lambda_b \), \( \alpha^a_n(\cdot) = \xi_n \alpha^a(\cdot) \) and \( \alpha^b_n(\cdot) = \xi_n \alpha^b(\cdot) \), while \( p_{n, a}(q) = p^{\alpha^a_n}(q) / \xi_n \) and \( p_{n, b}(q) = p^{\alpha^b_n}(q) / \xi_n \).

Under Assumption 1 to 7, the corresponding price per-trade process \((\bar{a}_n(\cdot), \bar{b}_n(\cdot))\) converges weakly in \( D[0, \infty) \) to a pure jump process \((\bar{a}(\cdot), \bar{b}(\cdot))\) with jump rate at time \( t \) equal to

\[
\lambda(t) := \int \left( \lambda^a p^{\alpha^a}(\bar{a}(t-), \bar{b}(t-), q) + \lambda^b p^{\alpha^b}(\bar{a}(t-), \bar{b}(t-), q) \right) \pi(dq; \bar{a}(t-), \bar{b}(t-)).
\]

Given a jump occurs at time \( t \), its jump size follows the following distribution:

\[
P(\Delta \bar{a} = i, \Delta \bar{b} = j) = \lambda(t)^{-1} \int I(a(q) - \bar{a}(t-)) = i, b(q) - \bar{b}(t-) = j \\
\cdot \left( \lambda^a p^{\alpha^a}(\bar{a}(t-), \bar{b}(t-), q) + \lambda^b p^{\alpha^b}(\bar{a}(t-), \bar{b}(t-), q) \right) \pi(dq; \bar{a}(t-), \bar{b}(t-)).
\]

The proof of this result is given in Section 8.2. We introduced the upper bound \( \Gamma \) on the prices \( \bar{a}_n(\cdot) \) in order to facilitate the verification of the so-called compact containment condition in order to apply the averaging principle.
6 Longer Time Scales: Continuous Time Model

In previous sections, we have built a connection between the distribution of the price change per-trade and the dynamics of the order flow. In this section we will characterize the dynamics of the trading price in a longer time scale, as we will show. Our goal here is to explain how jumps in the continuous time price dynamics for the bid/ask price are related to the structure of the LOB.

Recall that $\bar{a}(t)$ and $\bar{b}(t)$ are the bid/ask price per-trade. Define $\bar{s}(t) = \bar{a}(t) - \bar{b}(t)$ to be the bid-ask spread per-trade at time $t$, and $\bar{m}(t) = (\bar{a}(t) + \bar{b}(t))/2$ the mid-price. We shall develop a stochastic model for the price-spread dynamics in longer time scale (order of several minutes or more). The model will be a jump-diffusion limit of the discrete price-spread processes as given in Section 2.

Consider a sequence of limit order books indexed by $n$ and their bid/ask price per-trade process $\{(\bar{a}^n(\cdot), \bar{b}^n(\cdot))\}$. The dynamic of $\bar{a}^n(\cdot), \bar{b}^n(\cdot))$ is characterized by the arrival rate of market orders and the price changes. By Theorem 1, the price change distribution at trade times can be put in a one-to-one correspondence with the sequence of traffic intensity (i.e. ratios between cancellations and arrival times at each relative price in the LOB). Because of this one-to-one correspondence, we shall directly work with price change distribution at the time-scale of market orders.

To simplicity of the exposition (as we are mainly concerned to show how jumps in continuous time process observed in the order of minutes are connected to the structure of the LOB), we shall assume that the price changes of the ask and bid sides follows the same distribution which depends only on the spread size. The proofs will be almost the same for the more general case where the ask and bid price changes have different distributions. Let $\{t_k\}$ be the sequence of the trade times. We use $\Delta^n_a(s^n(t_k))$ to denote the ask price change and $\Delta^n_b(s^n(t_k))$ the bid price change. For simplicity in the notation, we often write $\Delta^n_a(s^n(t_k))$ instead of $\Delta^n_a(s^n(t_k))$ (similarly for $\Delta^n_b(s^n(t_k))$). As $\Delta^n_a(t_k)$ and $\Delta^n_b(t_k)$ follow the same distribution, we simply provide the description for $\Delta^n_a(t_k)$ in our following Assumption 1.

Assumption 1. (Price change distribution) First define,

$$\Delta^n_a(s^n(t_k)) = (1 - I^a_k) \cdot (-1)^{R^u_k} [U^a_k/\sqrt{n\delta}] \delta + I^a_k[s^n(t_k)V^a_k/(2\delta)]\delta \quad (6)$$

where:

i) $I^a_k$ is Bernoulli with $P(I^a_k = 1) = q$ for some $q > 0$,

ii) $U^a_k$ is a random variable with support on $[0, \xi]$ for $\xi \in (0, \infty)$.

iii) $R^u_k$ is Bernoulli with $P(R^u_k = 1) = (1 + 2\beta/\sqrt{m})/2$ for some $\beta > 0$,

iv) $V^a_k$ is a continuous random variable so that $P(V^a_k \geq -1) = 1$.

v) the random variables $I^a_k, U^a_k, R^u_k$ and $V^a_k$ are independent of each other (independence is assumed to hold across $k$ and for the superindices $a,b$).

Then we let

$$\begin{align*}
a^n(t_{k+1}) &= a^n(t_k) + \Delta^n_a(s^n(t_k)) \lor ([-s^n(t_k)/(2\delta)] \delta), \\
b^n(t_{k+1}) &= b^n(t_k) - \Delta^n_b(s^n(t_k)) \lor ([-s^n(t_k)/(2\delta)] \delta),
\end{align*}$$

and this is equivalent to assuming

$$\theta(i\delta, a^n(t_k), b^n(t_k)) = P(\Delta^n_a(s^n(t_k)) \lor ([-s^n(t_k)/(2\delta)] \delta) \geq i\delta).$$

Remarks:

1. Recall that we have ruled out orders which cross over the mid-price, this results in the cap $\lor ([-s^n(t_k)/(2\delta)] \delta)$ appearing in (7), which consequently yields $a^n, b^n$. 

23
Theorem 3. For the $n$-th system, let $\bar{s}(t) = \bar{a}(t) - \bar{b}(t)$ be the spread process and $\bar{m}(t) = \bar{a}(t) + \bar{b}(t)$ be twice of the mid-price. Suppose $(\bar{s}(0), \bar{m}(0)) = (s_0, m_0)$. Then, under Assumption 2 and 3, the pair of processes $(\bar{s}(t), \bar{m}(t)) \in D([0, \infty), \mathbb{R}^+ \times \mathbb{R})$ converges weakly to $(\bar{s}, \bar{m}) \in D([0, \infty), \mathbb{R}^+ \times \mathbb{R})$ with $(\bar{s}(0), \bar{m}(0)) = (s_0, m_0)$ such that

$$
\begin{cases}
    d\bar{s}(t) = -\eta(t)dt + dW_\alpha(t) + dW_\beta(t) + \bar{s}(t_-)dJ_1(t)/2 + \bar{s}(t_-)dJ_2(t)/2 + dL(t), \\
    d\bar{m}(t) = dW_\alpha(t) - dW_\beta(t) + \bar{s}(t_-)dJ_1(t)/2 - \bar{s}(t_-)dJ_2(t)/2.
\end{cases}
$$

(8)
Here,

1. Let \( \bar{\mu}(t) = \frac{2\kappa \mu}{\kappa - \delta} + \left( 2\mu_0 - \frac{2\kappa \mu}{\kappa - \delta} \right) e^{-(\kappa-\delta)t} \).

2. Then, \( \eta(t) = -2\beta \left( \mathbb{E}[U^a_k] + \mathbb{E}[U^b_k] \right) \cdot \bar{\mu}(t) \)

3. \( W_a \) and \( W_b \) are two diffusion processes, such that

\[
W_a(t) = \sqrt{\mathbb{E}[(U^a_k)^2]} \cdot B_1 \left( \int_0^t \bar{\mu}(s) ds \right), \quad W_b(t) = \sqrt{\mathbb{E}[(U^b_k)^2]} \cdot B_2 \left( \int_0^t \bar{\mu}(s) ds \right)
\]

where \( B_1 \) and \( B_2 \) are two independent Brownian motions.

4. \( J_1 \) and \( J_2 \) are two i.i.d. compound non-homogeneous Poisson processes with time-varying jump intensity \( \gamma \bar{\mu}(t) \) and the jump density distribution given by the density of \( V^a_1 \).

5. \( s(t) \geq 0 \) and \( L(t) \) satisfies: \( L(t) = 0, \ dL(t) \geq 0 \) and \( s(t)dL(t) = 0 \) for all \( t \geq 0 \).

Theorem 3 underscores the role that the LOB structure plays in the price dynamics over the course of a trading day. In particular, it is remarkable that the LOB appears to inform price-jump dynamics at time scales of the order of several minutes up to hours.

7 Conclusions and additional potential extensions

In this paper we develop a Markovian model for the limit order book of small-tick stocks. We propose an asymptotic regime under which we derive an analytic expression of the price change process in terms of the arrival and cancellation rates of the limit orders. These analytical expressions are justified using multi-scale analysis and the stochastic-averaging framework. We have explored extensions which include dependence between arriving market and limit orders in Section 5.

Our multiscale approach be further extended without much complications to include other types of dependence between dynamics in the arrivals of the market orders. For instance, one way to extend our model is to allow traders to post market orders depending on the current bid/ask price. This modification can be introduced by thinning the original Hawkes arrivals of the market orders, the thinning parameter might depend on the observed bid/ask price \((a(\cdot), b(\cdot))\) to reflect traders’ reaction on market conditions.

Other examples of the interactions between market participants that can be included in our model extensions are correlation between the bid and ask sides and dependence between arrival rate of market orders and the spread width.

8 Appendix: Technical Proofs

8.1 The Proof of Theorem 1

Proof of Theorem 1. In this proof, we will follow the notations required to apply Theorem 2.1 and Example 2.3 in Kurtz [17]. Define

\[
X_n(\cdot) = \left( \bar{a}_n(\cdot), \bar{b}_n(\cdot), \mu^a(\cdot), \mu^b(\cdot) \right) \in E_1 := \mathbb{R}^4.
\]
In addition, let $Y_n(\cdot) = (Y^n_i(\cdot) : i \geq 1) \in \mathbb{Z}^\infty$, where $Y_n = q_n$ encodes the number of limit orders as explained in Section 2.1 (i.e. the number of orders at level $i\delta$ is encoded with a negative sign if these are buy orders and with a positive sign if they are sell orders). We use $E_2$ to denote the space $\mathbb{Z}^\infty$ endowed with the topology induced by the metric

$$d(y, z) = \sum_{i=1}^{\infty} 2^{-i} \min\left(\left| y^i - z^i \right|, 1 \right),$$

which is equivalent to the pointwise convergence topology (i.e. $\{y(n)\} \subset E_2$ converges to $y$ if $y_k(n) \to y_k$ for any given $k \geq 1$). Because $y^i$ is an integer, the topology induced is equivalent to the discrete topology on finitely many coordinates. We naturally endow $E_1$ with the Euclidean metric.

Then, $\{(X_n(\cdot), Y_n(\cdot))\}$ is a sequence of stochastic processes living in the product space $E_1 \times E_2$. To be precise, let us explicitly describe the dynamics of this Markov chain.

Equation (1) describes the dynamics of $(\mu^a(\cdot), \mu^b(\cdot))$. Actually, if $\tilde{N}^a(\cdot)$ and $\tilde{N}^b(\cdot)$ are two independent Poisson processes with unit rate, we can couple the dynamics (1) with the explicit representations

$$N^a(t) = \tilde{N}^a \left( \int_0^t \mu^a(s) ds \right) \quad \text{and} \quad N^b(t) = \tilde{N}^b \left( \int_0^t \mu^b(s) ds \right).$$

Using this representation, we can redefine $t_k$ as follows. Set $t_0 = 0$, and for $k \geq 1$, let

$$t_k = \inf\{ t > t_{k-1} : N^a(t) - N^a(t_-) + N^b(t) - N^b(t_-) \neq 0 \}.$$

For each $i$, let $\{\tilde{N}^a_{i,p}(\cdot)\}_{i \geq 1}$, $\{\tilde{N}^b_{i,p}(\cdot)\}_{i \geq 1}$, $\{\tilde{N}^a_{i,\alpha}(\cdot)\}_{i \geq 1}$, $\{\tilde{N}^b_{i,\alpha}(\cdot)\}_{i \geq 1}$ be four independent sequences of Poisson processes with unit rate and independent of $N^a(\cdot)$ and $N^b(\cdot)$. Then, define

$$N^a_{i,n}(t) = \tilde{N}^a_{i,p} \left( \int_0^t \xi_n \lambda^a i\delta; \bar{a}_n(s), \bar{b}_n(s) \right) ds,$$

$$N^b_{i,n}(t) = \tilde{N}^b_{i,p} \left( \int_0^t \xi_n \lambda^b i\delta; \bar{a}_n(s), \bar{b}_n(s) \right) ds,$$

$$N^a_{i,n}(t) = \tilde{N}^a_{i,\alpha} \left( \int_0^t \alpha^a i\delta; \bar{a}_n(s), \bar{b}_n(s) \right) ds,$$

$$N^b_{i,n}(t) = \tilde{N}^b_{i,\alpha} \left( \int_0^t \alpha^b i\delta; \bar{a}_n(s), \bar{b}_n(s) \right) ds,$$

representing the number of arrivals of limit orders and the number of canceled limit orders at the relative price level $i\delta$ up to time $t$ in the $n$-th LOB system. Therefore, $Y_n(\cdot)$, defined as the number of limit orders waiting at price level $i\delta$ in the $n$-th LOB system at current time, changes at time $t$ by

$$dY_n^{i\delta; \bar{a}_n(t)}(t) = dN^a_{i,n}(t) - dN^a_{i,n}(t) \quad \text{if } i\delta \geq \frac{\bar{a}_n(t) + \bar{b}_n(t)}{2} \quad \text{and otherwise } dY_n^{b_n(t); i\delta}(t) = dN^b_{i,n}(t) - dN^b_{i,n}(t),$$

and finally note that

$$\bar{d}_n(\cdot) = a_n \left( t_{(N^a + N^b)(\cdot)} \right) \quad \text{and} \quad \bar{b}_n(\cdot) = b_n \left( t_{(N^a + N^b)(\cdot)} \right).$$

26
It is clear from the previous construction that for each \( n \), \( (X_n(\cdot), Y_n(\cdot)) \) is a continuous time Markov Chain. The process is non-explosive because \( \sum_i p_a^h(i\bar{\delta};\bar{a}_n(s), \bar{b}_n(s)) = 1 \).

Now, let us introduce a few definitions. Let \( C(E_1) \) be the space of all continuous functions \( f : E_1 \to R \) that vanish at infinity and let \( C(E_1) \) be the space of all bounded continuous functions \( f : E_1 \to R \).

Next, define \( D \) to be the set of function \( h : E_1 \times E_2 \to R \) satisfying that for each \( y \in E_2 \),

\[
h(\cdot, y) \in C(E_1), \quad \partial h(\cdot, y) / \partial x_3 \in C(E_1), \quad \partial h(\cdot, y) / \partial x_4 \in C(E_1).
\]

Now, we define \( A : D \to C(E_1 \times E_2) \) as

\[
Ah(x, y) = \mu^a \cdot [h(a(y), b(y), \mu^a + \delta_1, \mu^b + \delta_2, y) - h(x, y)]
+ \mu^b \cdot [h(a(y), b(y), \mu^a + \delta_2, \mu^b + \delta_1, y) - h(x, y)]
- \kappa (\mu^a - \mu) \cdot \partial h / \partial x_3 (x, y) - \kappa (\mu^b - \mu) \cdot \partial h / \partial x_4 (x, y).
\]

and \( B : D \to C(E_1 \times E_2) \) as

\[
Bh(x, y) = \sum_i \{ \lambda^a p^a(i\bar{\delta};\bar{a}, \bar{b}) h(x, y + e_{\bar{a}+i}) + \lambda^b p^b(i\bar{\delta};\bar{a}, \bar{b}) h(x, y - e_{\bar{b}-i})
- \left( \lambda^a + \lambda^b \right) h(x, y) \}
+ \sum_i \{ \alpha^a(i\bar{\delta};\bar{a}, \bar{b}) y^{a+\bar{i}} ((h(x, y - e_{\bar{a}+i}) - h(x, y))
+ \alpha^b(i\bar{\delta};\bar{a}, \bar{b}) y^{b-\bar{i}} (h(x, y + e_{\bar{b}-i}) - h(x, y)) \}.
\]

From the stochastic representation provided earlier it is straightforward to verify that the generator of \( (X_n(\cdot), Y_n(\cdot)) \) is given by

\[
Qh(x, y) \triangleq Ah(x, y) + \xi_n Bh(x, y),
\]

and therefore

\[
M^h_n(t) \triangleq h(X_n(t), Y_n(t)) - \int_0^t Qh(X_n(s), Y_n(s)) \, ds
\]

is a local martingale.

Next, pick \( h \in D \) and suppose that \( h(x, y) = f(x) \) for \( f : E_1 \to R \) (i.e. \( h(x, y) \) depends only on \( x \)). Then, we have that

\[
Qh(x, y) = Ah(x, y)
\]

and \( M^h_n(\cdot) \) is a martingale.

Now, pick \( h \in D \) such that \( h(x, y) = g(y^1, y^2, ..., y^k) \) for any fixed \( k < \infty \), then

\[
Qh(x, y) = \xi_n Bh(x, y)
\]

and \( M^h_n(\cdot) \) is also a martingale.

In order to apply the Theorem 2.1 (stochastic averaging principle) in Kurtz [17], we need to show (1) that \( \{ Y_n(t) : t \geq 0, n = 1, 2, ... \} \) are tight in \( E_2 \); (2) \( X_n(\cdot) \) satisfies the compact containment condition that is, we need to show that for any \( T > 0 \) and \( \epsilon > 0 \), there exists a compact set \( K \subset E_1 \) such that

\[
\lim_{n \to \infty} P \{ X_n(t) \in K \text{ for all } t \in [0, T] \} \geq 1 - \epsilon.
\]
We first consider $\mu^a(T)$ and $\mu^b(T)$. By [13] with the assumption that $\mu > 0$, $\kappa > 0$, $\delta_1 > \delta_2 > 0$ and $\delta_1 + \delta_2 < \kappa$, we have the that the Hawkes process is non-explosive (i.e. is bounded almost surely on any compact time interval), since the spectral radius of the matrix is

$$M = \begin{bmatrix} \delta_1/\kappa & \delta_2/\kappa \\ \delta_2/\kappa & \delta_1/\kappa \end{bmatrix}.$$

Next, to ensure (9) we need to deal with $\bar{a}_n(\cdot)$ and $\bar{b}_n(\cdot)$, which simply are piecewise constant processes changing values at the times $0 < t_1 < t_2 < \ldots < t_{N^a(T)+N^b(T)}$ and such that $\bar{a}_n(t_k) = a_n(t_k)$ and $\bar{b}_n(t_k) = b_n(t_k)$.

Because of the compact containment condition established for $\mu^a(\cdot)$ and $\mu^b(\cdot)$ we know that $N = N^a(T) + N^b(T)$ is finite almost surely. We can condition on

$$\mathcal{G}_T = \sigma\left\{\mu^a(s), \mu^b(s), N^a(s), N^b(s) : 0 \leq s \leq T\right\},$$

and proceed sequentially for $k = 1, 2, \ldots, N$, verifying that $a_n(t_k)$ and $b_n(t_k)$ are tight. We first consider $k = 1$, i.e. $s \in [0,t_1)$. Note that for each fixed $i$, we can represent $|Y^i_n(s)| = |Y^i(\xi_n s)|$, where $|Y^i(s)|$ is simply the number of customers-in-system in an infinite-server queue at time $s$.

The arrival rate of customers in this system is bounded from above by $\lambda^a + \lambda^b$ and the service rate is bounded from below by $\max\{a^i(\bar{a}_n(0); \bar{a}(0), b(0)), a^b(b_n(0) - \bar{a}; \bar{a}(0), b(0))\} > 0$. Therefore, $|Y^i_n(s)|$ converges weakly as $n \to \infty$ to a Poisson random variable. Actually, we have that

$$Y^i_n(s) \Rightarrow Y^i = I\left(1 - \frac{1}{\mu^a(0)}\right) Z^{i,a} - I\left(1 - \frac{1}{\mu^b(0)}\right) Z^{i,b},$$

where

$$Z^{i,a} \sim \text{Poisson}\left\{\frac{\lambda^a}{\mu^a} \frac{a^a((i - \bar{a}(0)) \delta; \bar{a}(0), b(0))}{\alpha^a((i - \bar{a}(0)) \delta; \bar{a}(0), b(0))}\right\},$$

$$Z^{i,b} \sim \text{Poisson}\left\{\frac{\lambda^b}{\mu^b} \frac{a^b((b(0) - i) \delta; \bar{a}(0), b(0))}{\alpha^b((b(0) - i) \delta; \bar{a}(0), b(0))}\right\}.$$  \hfill (11)

By the thinning theorem the sequence $\{Y^i_n(s)\}_{i \geq 1}$ is conditionally independent given $\mathcal{G}_{t_i}$ and so weak convergence of $\{Y^i_n(s)\}_{i \geq 1}$ towards $\{Y^i\}_{i \geq 1}$ as $n \to \infty$ occurs under the discrete topology on finitely many coordinates (the coordinates of $Y = (Y^1, Y^2, \ldots)$ are conditionally independent with limiting distribution (10)).

Because of (2), the functions $a(\cdot)$ and $b(\cdot)$ are continuous under the discrete topology on finitely many coordinates of $y$ on the support of the random element $\{Y(s)\}_{i \geq 1}$. So, we have that

$$(a(Y_n(t_1)), b(Y_n(t_1))) = (\bar{a}_n(t_1), \bar{b}_n(t_1)) \Rightarrow (\bar{a}(t_1), \bar{b}(t_1))$$ \hfill (12)

as $n \to \infty$.

So, from the argument leading to (12), applied subsequently on the intervals $[t_1, t_2], [t_{N-1}, t_N]$, we can verify two conditions required in the application of Theorem 2.1 from Kurtz [17].

Finally, we need to check that for all $T > 0$ and $h \in \mathcal{D}$ with $h(x, y) = f(x)$ for some $f : E_1 \to R$,

$$\int_0^T E \left(|Ah(X_n(s), Y_n(s))|^2\right) ds < \infty.$$
By Gronwall’s lemma we conclude that (13) holds.

\[
\sup_{s \in [0, T]} E \left( \mu^a(s)^2 + \mu^b(s)^2 \right) < \infty. \tag{13}
\]

Note from (1) that there is a continuous positive function \( r(\cdot) \) and a constant \( c \in (0, \infty) \) such that

\[
E \left( \mu^a(t)^2 + \mu^a(t)^2 \right) \leq r(T) + c \int_0^T E \left( \mu^a(s)^2 + \mu^a(s)^2 \right) ds.
\]

By Gronwall’s lemma we conclude that (13) holds.

Then, following Theorem 2.1 and Example 2.3 in Kurtz [17] we obtain that \( \{X_n(\cdot)\} \) is tight in \( D([0, \infty), E_1) \) and each limit process along subsequences, \( X(\cdot) = (\bar{a}(\cdot), \bar{b}(\cdot), \mu^a(\cdot), \mu^b(\cdot)) \), satisfies that for \( f \in C(E_1) \) such that \( h_f(x, y) := f(x) \in D \), the stochastic process

\[
f(X(t)) - \int_0^t \int_{E_2} Ah_f(X(s), y) \pi_{X(s)}(dy) ds,
\]

is a martingale, where \( \pi_{X(s)}(\cdot) \) is the unique stationary distribution of a stochastic process \( Y \in E_2 \) which solves the martingale problem

\[
g(Y(t)) - \int_0^t B h_g(x, Y(u)) du.
\]

for \( h_g(x, y) := g(y) \in D \). In our case, \( \pi_x(\cdot) \) is simply the distribution of the sequence of independent random variables \( \{Y^i\}_{i \geq 1} \) given \( (\bar{a}(0), \bar{b}(0)) \). Now we compute that in (14),

\[
\int_{E_2} Af(X(s), y) \pi_{X(s)}(dy)
\]

\[
= \sum_{i,j} \left\{ \mu^a(s) \cdot \left[ f(\bar{a}(s) + i \delta, \bar{b}(s) - j \delta, \mu^a(s) + \delta_1, \mu^b(s) + \delta_2) - f(\bar{a}(s), \bar{b}(s), \mu^a(s), \mu^b(s)) \right] \\
+ \mu^b(s) \cdot \left[ f(\bar{a}(s) + i \delta, \bar{b}(s) - j \delta, \mu^a(s) + \delta_1, \mu^b(s) + \delta_2) - f(\bar{a}(s), \bar{b}(s), \mu^a(s), \mu^b(s)) \right] \\
- \kappa(\mu^a(s) - \mu) \cdot \frac{\partial f}{\partial x_3}(\bar{a}(s), \bar{b}(s), \mu^a(s), \mu^b(s)) \\
- \kappa(\mu^b(s) - \mu) \cdot \frac{\partial f}{\partial x_4}(\bar{a}(s), \bar{b}(s), \mu^a(s), \mu^b(s)) \right}\} \cdot \pi_{X(s)}(\{i, j\})
\]

\[
= \sum_{i,j} \left\{ \mu^a(s) \cdot \left[ f(\bar{a}(s) + i \delta, \bar{b}(s) - j \delta, \mu^a(s) + \delta_1, \mu^b(s) + \delta_2) - f(\bar{a}(s), \bar{b}(s), \mu^a(s), \mu^b(s)) \right] \\
+ \mu^b(s) \cdot \left[ f(\bar{a}(s) + i \delta, \bar{b}(s) - j \delta, \mu^a(s) + \delta_1, \mu^b(s) + \delta_2) - f(\bar{a}(s), \bar{b}(s), \mu^a(s), \mu^b(s)) \right] \\
\right\} \cdot \left[ \theta^a (i \delta; \bar{a}(s), \bar{b}(s)) - 3^a (i \delta; \bar{a}(s), \bar{b}(s)) \right] \times \left[ \theta^b (j \delta; \bar{a}(s), \bar{b}(s)) - 3^b (j \delta; \bar{a}(s), \bar{b}(s)) \right]
\]

One can check that the martingale problem (14) has a unique solution \( X(\cdot) \), see for instance Chapter 4.4 in Ethier and Kurtz [9]. In particular, \((\bar{a}(\cdot), \bar{b}(\cdot))\) is equivalent in distribution to a jump process with jump intensity \( \mu(s) = \mu^a(s) + \mu^b(s) \) at time \( s \) and jump size distribution

\[
\Pr (\bar{a}(t) - \bar{a}(t-) = i, \bar{b}(t) - \bar{b}(t-) = -j | \bar{a}(t), \bar{b}(t)) \\
= [\theta^a (i; \bar{a}(t), \bar{b}(t)) - 3^a (i + 1; \bar{a}(t), \bar{b}(t))] \times [\theta^b (j; \bar{a}(t), \bar{b}(t)) - 3^b (j + 1; \bar{a}(t), \bar{b}(t))].
\]
Since \( \{X_n(\cdot)\} \) is tight and each of its convergent subsequence admits the same limit \( X(\cdot) \), we can conclude that \( \{X_n(\cdot)\} \) weakly converges to \( X(\cdot) \) in \( D[0, \infty) \). ■

8.2 Proof of Theorem 2

Proof of Theorem 2. We outline the argument following a very similar approach to that explained in Theorem 1. Define \( X_n(\cdot) = (\tilde{a}_n(\cdot), \tilde{b}_n(\cdot)) \in \mathbb{R}^2 \) and \( Y_n = q_n \) encodes the number of limit sell orders in the \( n \)-th LOB system at time \( t \). We use \( E_2 = \mathbb{Z}^T \) endowed with the Euclidean topology, which is equivalent, because \( y^i \) is an integer, the discrete topology. We also endow \( E_1 \) with the Euclidean metric.

The space of functions \( \hat{C}(E_1), \hat{C}(E_2) \) and \( D \) are all defined similarly as in the proof of Theorem 1. Define \( \mathcal{A} : D \to C(E_1 \times E_2) \) as

\[
\mathcal{A}f(x, y) = \lambda^a \sum_{i=1}^{\Gamma} \left( 1 - \frac{p^m(a(x), b(y), y)}{\xi_n} \right) p^a(i\delta; x, y) h(x, y + e_{a+i})
+ \lambda^b \sum_{i=1}^{\Gamma} \left( 1 - \frac{p^m(b(x), b(y), y)}{\xi_n} \right) p^b(i\delta; x, y) h(x, y + e_{b+i})
- \lambda \sum_{i=1}^{\Gamma} \left( 1 - \frac{p^m(a(x), b(y), y)}{\xi_n} \right) p^e(i\delta; x, y) h(x, y + e_{e+i})
\]

and \( \mathcal{B}_n : D \to C(E_1 \times E_2) \) as

\[
\mathcal{B}_n h(x, y) = \sum_{i=1}^{\Gamma} \left( \lambda^a \left( 1 - \frac{p^m(a(x), b(y), y)}{\xi_n} \right) p^a(i\delta; x, y) h(x, y + e_{a+i})
+ \lambda^b \left( 1 - \frac{p^m(b(x), b(y), y)}{\xi_n} \right) p^b(i\delta; x, y) h(x, y + e_{b+i})
- \lambda \left( 1 - \frac{p^m(a(x), b(y), y)}{\xi_n} \right) p^e(i\delta; x, y) h(x, y + e_{e+i}) \right),
\]

The process \( (X_n(\cdot), Y_n(\cdot)) \) is a Markov jump process with generator \( \mathcal{Q} \) such that for each \( h \in D \),

\[
\mathcal{Q} h(x, y) = \mathcal{A} h(x, y) + \xi_n \mathcal{B}_n h(x, y).
\]

Now for \( h_f \in D \) and suppose that \( h_f(x, y) = f(x) \) for \( f : E_1 \to R \) (i.e. \( h_f(x, y) \) depends only on \( x \)). we have that

\[
\mathcal{Q} h_f(X_n(s), y) = \mathcal{A} h_f(X_n(s), y).
\]

By the boundedness assumption on \( D \), we have that

\[
f(X_n(t)) - \int_0^t \mathcal{A} h_f(X_n(s), Y_n(s)) \, ds
\]

is a martingale. Likewise, we also have that if \( h_g(x, y) = g(y^1, ..., y^T) \) with \( h_g \in D \) then

\[
g(Y_n(t)) - \xi_n \int_0^t \mathcal{B}_n h_g(X_n(s), Y_n(s)) \, ds
\]

is a martingale.
Finally, since a sub-subsequence that converges weakly to some sub-limit process \( X_\pi \) in the sense that the stochastic process defined by (15) is a martingale. Moreover, in the expression (15) is the unique solution with jump intensity \( s \).

Following Chapter 4.4 in Ethier and Kurtz [9] one can verify that the martingale problem (15) is the unique stationary distribution of a stochastic process \( Y \in E_2 \) which satisfies the martingale problem

\[
\mathbb{E}h(x, y) = \sum_{i=1}^\Gamma \left[ \lambda^a p^a(i\delta; x, y)g(y) + \alpha^a (i\delta; x, y) \cdot y^{a+i} \cdot \left( g(y - e_{a+i}) - g(y) \right) + \alpha^b(i\delta; x, y) \cdot y^{b-i} \cdot \left( g(y - e_{b-i}) - g(y) \right) \right].
\]

Because \( \{X_n(\cdot)\} \) are jump processes with uniform bound \((\bar{a}(\cdot), \bar{b}(\cdot) \in [0, \Gamma \delta])\), and hence are tight on \( E_1 \). Besides, under Assumption 7, \( \{Y_n(t) : t > 0, n = 1, 2, \ldots\} \) are relatively compact. Finally, since \( k \) is bounded, for any \( T > 0 \), \( \int_0^T \mathbb{E} \left( |Ah(X_n(s), Y_n(s))|^2 \right) ds < \infty \). Then according to Theorem 2.1 and the subsequent Example 2.3 in Kurtz [17], every subsequence of \( \{X_n(\cdot)\} \) admit a sub-subsequence that converges weakly to some sub-limit process \( X(\cdot) \), and each sub-limit process \( X(\cdot) = (\bar{a}(\cdot), \bar{b}(\cdot), \lambda^a(\cdot), \lambda^b(\cdot)) \) is a solution to the martingale problem:

\[
f(X(t)) - \int_0^t \int_{E_2} Ahf(X(s), y) \pi_X(dy)ds, \tag{15}
\]

in the sense that the stochastic process defined by (15) is a martingale. Moreover, in the expression (15), \( \pi_x(\cdot) \) is the unique stationary distribution of a stochastic process \( Y \in E_2 \) which satisfies the martingale problem

\[
g(Y(t)) - \int_0^t \mathbb{E}h(x, Y(u))du.
\]

In our case, \( \pi_x(\cdot) \) is simply the unique stationary distribution \( \pi(\mathbf{q}; a(t_k), b(t_k)) \) under the arrival rates \( \lambda^a p^a(i\delta; a(t_k), b(t_k), \mathbf{q}) \) and \( \lambda^b p^b(i\delta; a(t_k), b(t_k), \mathbf{q}) \), and the cancellation rates \( \alpha^a(i\delta; a(t_k), b(t_k), \mathbf{q}) \) and \( \alpha^b(i\delta; a(t_k), b(t_k), \mathbf{q}) \). Now we compute that in (15),

\[
\int_{E_2} Ahf(X(s), y) \pi_X(dy) = \int_{E_2} \left\{ \lambda^a p^{m,a}(\bar{a}(s), \bar{b}(s), y) \cdot \left[ f(a(y), b(y)) - f(\bar{a}(s), \bar{b}(s)) \right] 
+ \lambda^b p^{m,b}(\bar{a}(s), \bar{b}(s), y) \cdot \left[ f(a(y), b(y)) - f(\bar{a}(s), \bar{b}(s)) \right] \right\} \cdot \pi(\mathbf{q}; \bar{a}(s), \bar{b}(s)).
\]

Following Chapter 4.4 in Ethier and Kurtz [9] one can verify that the martingale problem (15) has a unique solution \( X(\cdot) \). In particular, \( (\bar{a}(\cdot), \bar{b}(\cdot)) \) is equivalent in distribution to a jump process with jump intensity

\[
\lambda(s) = \lambda^a \int_{E_2} p^{m,a}(\bar{a}(s), \bar{b}(s), y) \pi_X(dy) + \lambda^b \int_{E_2} p^{m,b}(\bar{a}(s), \bar{b}(s), y) \pi_X(dy),
\]

at time \( s \), and jump size distribution at time \( s \) equal to

\[
P(\Delta \bar{a}(s) = i, \Delta \bar{b}(s) = j) = \lambda(s)^{-1} \int_{E_2} \mathbb{I}(a(y) - \bar{a}(s) = i\delta, b(y) - \bar{b}(s) = j\delta) \cdot \left( \lambda^a p^{m,a}(\bar{a}(s), \bar{b}(s), y) + \lambda^b p^{m,b}(\bar{a}(s), \bar{b}(s), y) \right) \pi_X(dy).
\]
Since \( \{X_n(\cdot)\} \) is tight and each of its convergent subsequence admits the same limit \( X(\cdot) \), we can conclude that \( \{X_n(\cdot)\} \) weakly converges to \( X(\cdot) \). ■

8.3 Model Calibration

**Treatment of hidden orders:** In our model, we define the bid/ask price per trade and use them as the benchmark price to determine the relative prices of limit orders between two consequent trades. If all trades are against visible orders, we can simply take the bid and ask prices right before the trade as the bid/ask price per trade, and the benchmark prices until the next trade. However, in real market, not all limit orders are displayed on the LOB and trades can happen against the so-called hidden limit orders. In our observation window, on average 20.53%, 22.08% and 16.62% of trades are against hidden limit orders for AMZN, BIDU and GS respectively. Those trades usually occur at a better price than the displayed bid/ask price and reveals the existence of possible hidden liquidity at the trade price to the public. In this light, if a trade is against some hidden order at the bid (ask) side, we take the trade price as the benchmark price until the next trade. For the other side, we still take the visible ask (bid) price at the time of trade as the benchmark price.

On the other hand, the distribution of the price change per-trade is computed based on the submission and cancellation rates of limit orders in our model. From the data, however, we can only calibrate the submission and cancellation rates of the visible limit orders. Therefore, we shall compute the empirical price change per trade between two consequent trades as the difference between the bid/ask price of the visible limit orders, right before the next trade, and the benchmark price.

After we determine the benchmark prices over all inter-trade intervals as described previously, we first transfer the dollar prices of all limit order events into relative prices with respect to the corresponding benchmark prices. For each trading day, we estimate the arrival and cancellation rates of the LOB model as follows.

Let \( L^\text{ask} \) be the number of sell limit orders that arrive during the trading day, and let \( D \) be the duration of the trading day which, as mentioned earlier, corresponds to 5.5 hours. Similarly, we use \( L^\text{bid} \) to denote the number of buy limit orders during the trading day. The arrival rates of the buy and sell limit orders, \( \lambda^b \) and \( \lambda^a \), are estimated as follows:

\[
\hat{\lambda}^b = \frac{L^\text{bid}}{D}, \quad \hat{\lambda}^a = \frac{L^\text{ask}}{D}.
\]

We then write \( L^\text{bid}_{\delta} \) (\( L^\text{ask}_{\delta} \)) to denote the total number of buy (sell) limit orders at relative price \( i\delta \) during the trading day. We estimate the probability of a limit order sitting at a relative price \( i\delta \), i.e. \( p^b(a; i\delta; a, b) \), at ask side and bid side to be

\[
\hat{p}^b(i\delta; a, b) = \frac{L^\text{bid}_{\delta}}{L^\text{bid}}, \quad \hat{p}^a(i\delta; a, b) = \frac{L^\text{ask}_{\delta}}{L^\text{ask}},
\]

respectively.

Next we use \( N^\text{bid} \) (\( N^\text{ask} \)) to denote the total number of buy (sell) market orders arriving during the trading day. Moreover, we use \( N^\text{bid}_{\geq \delta} \) (\( N^\text{ask}_{\geq \delta} \)) to denote the total number of market buy (sell) orders during the trading day, that arrive at a time at which the real-time bid (ask) price is greater than or equal to \( i\delta \) relatively to the benchmark price. The empirical estimates of tail probability of price change, i.e. \( \theta^b(i\delta; a, b) \) at each side are given by

\[
\hat{\theta}^b(i\delta; a, b) = \frac{N^\text{bid}_{\geq \delta}}{N^\text{bid}}, \quad \text{and} \quad \hat{\theta}^a(i\delta; a, b) = \frac{N^\text{ask}_{\geq \delta}}{N^\text{ask}}.
\]

32
Estimating the cancellation rate of limit orders, i.e. \( \alpha^b(i\delta; a, b), \alpha^a(i\delta; a, b) \), is not entirely straightforward as the frequency of cancellation events depends not only on the cancellation rate, but also on the queue size, or the volume of limit orders. The approach that we take is similar to the volume-dependent model introduced in [16].

For each relative price level \( i\delta \), either on the bid or ask side, we partition the limit order cancellation events into different bins, according to the queue sizes of limit orders on the price level, right before the events. The bins are labeled by integers \( j \geq 1 \), and the \( j \)-th bin contains all cancellation events right before which the queue sizes are in the set \( I_j = \{100 \cdot (j-1)+1, \cdots, 100 \cdot j\} \).

Our estimation of the cancellation involves several auxiliary quantities:

\[
C_{\text{bid/ask}}^j(i\delta) = \text{total size of canceled buy/sell limit orders in bin } j \text{ during the day},
\]
\[
T_{\text{bid/ask}}^j(i\delta) = \text{total time (secs) that buy/sell limit order queue sizes are in set } I_j \text{ during the day},
\]
\[
Q_{\text{bid/ask}}^j(i\delta) = 100 \cdot j - 50 = \text{mid value of set } I_j,
\]
\[
w_{\text{bid/ask}}^j(i\delta) = \frac{C_{\text{bid/ask}}^j(i\delta)}{\sum_{\text{bins } k} C_{\text{bid/ask}}^k(i\delta)}.
\]

The estimated cancellation rates are computed as

\[
\hat{\alpha}^b(i\delta; a, b) = \sum_{\text{bins } j \text{ at rel. price } i\delta} w_{\text{bid}}^j(i\delta) \times \frac{C_{\text{bid}}^j(i\delta)}{Q_{\text{bid}}^j(i\delta) \cdot T_{\text{bid}}^j(i\delta)},
\]
\[
\hat{\alpha}^a(i\delta; a, b) = \sum_{\text{bins } j \text{ at rel. price } i\delta} w_{\text{ask}}^j(i\delta) \times \frac{C_{\text{ask}}^j(i\delta)}{Q_{\text{ask}}^j(i\delta) \cdot T_{\text{ask}}^j(i\delta)}.
\]

We define \( 0 \cdot 0 = 0 \) in the previous equations. In order to maintain robustness in estimation, the bins with \( T_j(i\delta) \) of less than 0.05 milliseconds are considered empty.

The intuition behind our approach is that, under our model, the number of cancellation events in each bin \( j \), \( C_j(i\delta) \), should roughly equal to \( T_j(i\delta)Q_j(i\delta) \) multiplied by cancellation rate, given enough observations. Therefore, for each \( j \), the ratio \( C_j(i\delta)/(Q_j(i\delta) \cdot T_j(i\delta)) \) should give approximately equal estimation on the cancellation rate. However, bins containing smaller number of observations might give noisier estimates of the cancellation rate. In this light, we introduce the weights \( w_j(i\delta) \), that is proportional to the number of observations in bin \( j \), to help mitigate such noise.

### 8.4 The Proof of Theorem 3

Now let us describe the road map for the proof of Theorem 3. We first construct some auxiliary process \( \left( \tilde{S}^n(\cdot), \tilde{M}^n(\cdot) \right) \) living in the same probability space as the underlying process \( (\bar{s}^n(\cdot), \bar{m}^n(\cdot)) \). The auxiliary process is a convenient Markov process whose generator can be analyzed to conclude weak convergence to the postulated limiting jump diffusion (8). The auxiliary process has the same dynamics as the target process except when it is on the boundary-layer set \( [0, 2/\sqrt{n}] \times \mathbb{R} \). We also show the time spent by the two processes on the boundary-layer is small and as a result their difference caused by their different dynamics on the boundary is also small. Actually, such difference is negligible as \( n \to \infty \) and therefore the target process converges to the same limit process.
First, we define the auxiliary process coupled with the target process in a path by path fashion. Recall that by Assumption 6 and (7), we can write

\[
\begin{align*}
\hat{s}^n(t_{k+1}) &= \hat{s}^n(t_k) + \Delta^n_\alpha(\hat{s}^n(t_k)) \vee (-\hat{s}^n(t_k)/2\delta) + \hat{\nu}^n_\nu(\hat{s}^n(t_k)) \vee (-\hat{s}^n(t_k)/2\delta), \\
\hat{m}^n(t_{k+1}) &= \hat{m}^n(t_k) + \Delta^n_\alpha((\hat{s}^n(t_k))) \vee (-\hat{s}^n(t_k)/2\delta) - \hat{\nu}^n_\nu(\hat{s}^n(t_k)) \vee (-\hat{s}^n(t_k)/2\delta).
\end{align*}
\]

Now we define the auxiliary process \((\tilde{S}^n(\cdot), \tilde{M}^n(\cdot))\) coupled with \((\hat{s}^n(\cdot), \hat{m}^n(\cdot))\) as

\[
\begin{align*}
\tilde{S}^n(t_{k+1}) &= \tilde{S}^n(t_k) + \left(\Delta^n_\alpha(\tilde{S}^n(t_k)) + \Delta^n_\nu(\tilde{S}^n(t_k))\right) \vee (-\tilde{S}^n(t_k)), \\
\tilde{M}^n(t_{k+1}) &= \tilde{M}^n(t_k) + \Delta^n_\alpha((\tilde{S}^n(t_k))) - \Delta^n_\nu(\tilde{S}^n(t_k)),
\end{align*}
\]

(16)

with the initial condition \(\tilde{S}^n(0) = \hat{s}^n(0)\) and \(\tilde{M}^n(0) = \hat{m}^n(0)\).

Then the main result in this section Theorem 3 is an immediate corollary of the following two propositions.

**Proposition 1.** The auxiliary process \((\tilde{S}^n(\cdot), \tilde{M}^n(\cdot))\) converges weakly to the limit process given by (8).

**Proposition 2.** The difference process \((\hat{s}^n(\cdot) - \tilde{S}^n(\cdot), \hat{m}^n(\cdot) - \tilde{M}^n(\cdot))\) converges weakly to \((0, 0)\) on \(D_{\mathbb{R}^2}[0, t]\) for any \(t < \infty\).

**Proof of Proposition 1.** Define \(N_\alpha(t) = \sum_{j: t_j \leq t} I^n_j\), \(N_\beta(t) = \sum_{j: t_j \leq t} J^n_j\). Next, define \(\tilde{S}^n(0) = \hat{s}^n(0) \geq 0\), and \(\tilde{M}^n(0) = \hat{m}^n(0)\), and set

\[
S^n_\alpha(t) = \sum_{j: t_j \leq t} (-1)^{R^n_\alpha[U^n_j/(\sqrt{n}\delta_n)]} \delta_n, \quad S^n_\beta(t) = \sum_{j: t_j \leq t} (-1)^{R^n_\beta[U^n_k/(\sqrt{n}\delta_n)]} \delta_n.
\]

We will also define

\[
S^n(t) = S^n_\alpha(t) + S^n_\beta(t) + \hat{s}^n(0),
\]

\[
M^n(t) = S^n_\alpha(t) - S^n_\beta(t) + \hat{m}^n(0),
\]

(so by convention we set \(t_0 = 0\) and \(S^n(0) = \hat{s}^n(0)\)). Also, we define

\[
R^n_1(t) = S^n(t) - \min(S^n(u) : u \leq t, 0).
\]

Let \(A = \inf\{t \geq 0 : N_\alpha(t) \geq 1\}\) and \(B = \inf\{t \geq 0 : N_\beta(t) \geq 1\}\) be the first arrival times of \(N_\alpha(\cdot)\) and \(N_\beta(\cdot)\), respectively. Since

\[
\tilde{S}^n(t_k) = (\Delta^n_\alpha(\tilde{S}^n(t_k)) + \Delta^n_\nu(\tilde{S}^n(t_k)) + \hat{\nu}^n_\nu(\tilde{S}^n(t_k)))^+,
\]

we have that on \(\min(A, B) > t\)

\[
\tilde{S}^n(t) = R^n(t).
\]

The strategy proceeds as follows. Step 1: Show that if \((\hat{s}^n(0), \hat{m}^n(0)) \Rightarrow (\hat{s}(0), \hat{m}(0))\), the processes \((S^n(t), M^n(t) : t \geq 0)\) converges weakly in \(D[0, \infty)\) to the process \((X(t) : t \geq 0)\) defined via

\[
X_1(t) = \hat{s}(0) + \eta(t) + W_\alpha(t) + W_\beta(t),
\]

\[
X_2(t) = \hat{m}(0) + W_\alpha(t) - W_\beta(t).
\]

34
Step 2): Once Step 1) has been executed we can directly apply the continuous mapping principle to conclude joint weak convergence on \([0, \min(A, B)]\) of the processes

\[
R^n(\cdot) \Rightarrow R(\cdot) := X_1(\cdot) - \min (X_1(u) : 0 \leq u \leq \cdot),
\]

\[
M^n(\cdot) \Rightarrow X_2(\cdot).
\]

Step 3): By invoking the Skorokhod embedding theorem, we can assume that the joint weak convergence in Step 2) occurs almost surely. We can add the jump right at time \(\min(\cdot)\) (Step 3): By invoking the Skorokhod embedding theorem, we can assume that the joint weak convergence on \([0, \min(A, B)]\) of the processes

\[
D = I(A < B) R(A) V^a / 2 + I(B \leq A) R(B) V^b / 2,
\]

where \(V^b\) and \(V^a\) are i.i.d. copies of \(V_k^b\) and \(V_k^a\) respectively, and we also define

\[
D^n = I(A < B) [R^n(A) V^a / (2\delta)] \delta + I(B \leq A) [R^n(B) V^b / (2\delta)] \delta.
\]

Then put on \(t \in [0, \min(A, B)]\)

\[
\tilde{S}^n(t) = R^n(t) I(t < \min(A, B)) + I(t = \min(A, B)) (R^n(\min(A, B)) + D^n),
\]

\[
s(t) = R(t) I(t < \min(A, B)) + I(t = \min(A, B)) (R(\min(A, B)) + D),
\]

\[
\tilde{M}^n(t) = M^n(t) I(t < \min(A, B))
\]

\[
+ I(t = \min(A, B)) I(A < B) [R^n(A) V^a / (2\delta)] \delta
\]

\[
- I(t = \min(A, B)) I(B \leq A) [R^n(B) V^b / (2\delta)] \delta,
\]

\[
\tilde{m}(t) = X_2(t) I(t < \min(A, B))
\]

\[
+ I(t = \min(A, B)) I(A < B) R(A) V^a / 2\delta
\]

\[
- I(t = \min(A, B)) I(B \leq A) R(B) V^b / 2\delta.
\]

So, assuming Step 2) and using Skorokhod embedding we then conclude that

\[
\sup_{0 \leq t \leq \min(A, B)} \left| \tilde{S}^n(t) - \tilde{s}(t) \right| + \sup_{0 \leq t \leq \min(A, B)} \left| \tilde{M}^n(t) - \tilde{m}(t) \right| \to 0
\]

almost surely. Step 4): Finally, note that the convergence extends throughout the interval \([0, t]\) by repeatedly applying Steps 1) to 3) given that there are only finitely many jumps in \([0, t]\), which is guaranteed by tightness of processes \(N_a(t)\) and \(N_b(t)\) as the scaled limits of Hawkes processes with stationarity condition \(\delta / \kappa < 1\) imposed. Clearly then this procedure completes the construction to the solution of the SDE (8). So, we see that everything rests on the execution of Step 1), and for this we invoke the martingale central limit theorem (see Ethier and Kurtz [9], Theorem 7.1.4). Define

\[
Z_k^a(n) = (-1)^{R_k^a} \left[ U_k^a / (\sqrt{n}\delta_n) \right] \delta_n, \quad Z_k^b(n) = (-1)^{R_k^a} \left[ U_k^b / (\sqrt{n}\delta_n) \right] \delta_n.
\]

We have that

\[
\mathbb{E} Z_k^a(n) = \mathbb{E} Z_k^b(n) = (-2\beta / \sqrt{n}) \mathbb{E} \left( \left[ U_k^a / (\sqrt{n}\delta_n) \right] \delta_n \right) = -\frac{2\beta \mathbb{E} (U_k^a)}{n} + o(1/n)
\]

and

\[
\text{Var}(Z_k^a(n)) = \mathbb{E} \left( \left[ U_k^a / (\sqrt{n}\delta_n) \right]^2 \delta_n^2 \right) - O(1/n).
\]
Let $\mu^n_a(t)$ ($\mu^n_b(t)$) denote the intensity of market sell (buy) order at time $t$, and let $N^n_a(t)$ ($N^n_b(t)$) denote the counting process of market sell (buy) order arrivals in the $n$-th system. By the scaling specified in Assumption 3, we have

$$
\begin{align*}
    d\mu^n_a(t) &= -\kappa(\mu^n_a(t) - n\mu)dt + \delta_1 dN^n_a(t) + \delta_2 dN^n_b(t), \\
    d\mu^n_b(t) &= -\kappa(\mu^n_b(t) - n\mu)dt + \delta_1 dN^n_b(t) + \delta_2 dN^n_a(t),
\end{align*}
$$

and $\mu^n_a(0) = \mu^n_b(0) = n\mu_0 > 0$. Let $\mu_n(t) := \mu^n_a(t) + \mu^n_b(t)$ and $N_n(t) := N^n_a(t) + N^n_b(t)$, thus

$$
\begin{align*}
    \begin{cases}
        d\mu_n(t) &= -\kappa(\mu_n(t) - 2n\mu)dt + (\delta_1 + \delta_2)dN_n(t), \\
        \mu_n(0) &= 2n\mu_0
    \end{cases}
\end{align*}
$$

(17)

Let $\hat{\mu}_n(t) = \mu_n(t)/n$, and $\hat{\Lambda}_n(t) = \int_0^t \hat{\mu}_n(s)ds$. Since $Z^n_k(\cdot)$ is independent of $N_n(\cdot)$, we use standard the strong approximation results for random walks and renewal processes, see Chapter 2 of [8] and Theorem 2.1 c) in [14], to obtain

$$
\begin{align*}
    S^n_a(t) &= \sum_{k=1}^{N_n(t)} Z^n_k(n) = \bar{N}(t, \mu_n(s)ds) + \sum_{k=1}^{N_n(t)} Z^n_k(n) \\
    &= \mathbb{E}[Z^n_k(n)] \int_0^t \mu_n(s)ds + \sqrt{\text{Var}(Z^n_k(n))} B \left( \int_0^t \mu_n(s)ds \right) + o \left( \left( \int_0^t \mu_n(s)ds \right)^{1/2} \right) \\
    &= -2\beta \mathbb{E}[U^n_k] \cdot \hat{\Lambda}_n(t) + \frac{\sqrt{\mathbb{E}[U^n_k]}}{\sqrt{n}} B \left( n\hat{\Lambda}_n(t) \right) + o \left( \left( n\hat{\Lambda}_n(t) \right)^{1/2} \right)
\end{align*}
$$

Let $\delta = \delta_1 + \delta_2$, then

$$
\hat{\mu}_n(t) - 2\mu_0 = \int_0^t \kappa(2\mu - \hat{\mu}_n(s))ds + \delta \frac{N_n(t)}{n}
$$

By the law of large numbers, $N_n(\cdot)/n \Rightarrow E[N(\cdot)] = \int_0^t \hat{\mu}(s)ds$ on $D(0, \infty)$, where $N(t)$ is a Hawkes process with rate $\mu(t)$ such that

$$
\begin{align*}
    d\mu(t) &= -\kappa(\mu(t) - 2\mu)dt + \delta dN(t), \\
    \mu(0) &= 2\mu_0.
\end{align*}
$$

In particular, a standard convergence argument as in, for instance, Section 3.3 in Ethier and Kurtz [9] allows to conclude that $\bar{\mu}(t)$ satisfies the following equation,

$$
\bar{\mu}(t) - 2\mu_0 = \int_0^t \kappa(2\mu - \bar{\mu}(s))ds + \delta \int_0^t \bar{\mu}(s)ds.
$$

And by the continuous mapping theorem, $\bar{\mu}_n(\cdot)$ converges to $\bar{\mu}_\infty(\cdot)$, which is the solution to the ODE:

$$
\bar{\mu}_\infty(t) - 2\mu = \int_0^t \kappa(2\mu - \bar{\mu}_\infty(s))ds + \delta \int_0^t \bar{\mu}(s)ds,
$$

from which we solve $\bar{\mu}_\infty = \bar{\mu}$. Since we can solve the ODE

$$
\bar{\mu}(t) = \frac{2\kappa \mu}{\kappa - \delta} + \left( 2\mu_0 - \frac{2\kappa \mu}{\kappa - \delta} \right) e^{-(\kappa - \delta)t}.
$$
Applying the continuous mapping theorem again, we have as $n \to \infty$

$$\hat{\Lambda}_n(t) \Rightarrow \hat{\Lambda}_\infty(t) = \int_0^t \hat{\mu}(s)ds = \frac{2\kappa \mu t}{\kappa - \delta} - \frac{2\mu_0 - \frac{2\kappa \mu}{\kappa - \delta}}{\kappa - \delta} \left(e^{-(\kappa - \delta)t} - 1\right).$$

Thus

$$S_{a}^n(t) \Rightarrow -2\beta \mathbb{E}[U_{k}^a] \hat{\Lambda}_\infty(t) + \sqrt{\mathbb{E}[(U_{k}^a)^2]} \cdot B_1 \left(\hat{\Lambda}_\infty(t)\right)$$

under the Skorokhod topology on compact sets. A completely analogous strategy is applicable to conclude

$$S_{b}^n(t) \Rightarrow -2\beta \mathbb{E}[U_{k}^b] \hat{\Lambda}_\infty(t) + \sqrt{\mathbb{E}[(U_{k}^b)^2]} \cdot B_2 \left(\hat{\Lambda}_\infty(t)\right),$$

where $B_1(\cdot)$ and $B_2(\cdot)$ are two independent Brownian motions. Let

$$\eta(t) = -2\beta \left(\mathbb{E}[U_{k}^a] + \mathbb{E}[U_{k}^b]\right) \cdot \hat{\Lambda}_\infty(t),$$

$$W_a(t) = \sqrt{\mathbb{E}[(U_{k}^a)^2]} \cdot B_1 \left(\hat{\Lambda}_\infty(t)\right),$$

$$W_b(t) = \sqrt{\mathbb{E}[(U_{k}^b)^2]} \cdot B_2 \left(\hat{\Lambda}_\infty(t)\right)$$

then the convergence holds jointly due to independency of limit processes and therefore we obtain the conclusion required in Step 1). As indicated earlier, Steps 2) to 4) now follow directly. Note that now $J_1$ and $J_2$ are compound Poisson processes with jump intensity $\gamma \hat{\mu}(t).$ $\blacksquare$

**Proof of Proposition 2.** For simplicity, we assume that $\xi = 1$, otherwise we can divide $\tilde{S}$, $M$ and $\bar{s}$, $\bar{m}$ by the constant $\xi$. Assume also that $V_{k}^a \leq c$ for some $c \geq 1$. The general case can be dealt with using truncation because there are only a Poisson number of jumps that arise in $O(1)$ time. Now, let us first give a bound for the difference $\bar{s}^n(\cdot) - \tilde{S}^n(\cdot)$. For fixed $n$, we define $N(t) = \sum_{j: t_j \leq t} (I_{j}^a + I_{j}^b)$, intuitively $N(t)$ corresponds to the number of jumps of the limiting process from time 0 to time $t$. Now we prove by induction that

$$0 \leq \bar{s}^n(t_k) - \tilde{S}^n(t_k) \leq \left((1 + c)^{N(t_k)} - 1\right) \cdot \frac{2}{\sqrt{n}}. \quad (18)$$

At $t_0 = 0$, we have $\tilde{S}^n(t_0) = \bar{s}(t_0)$. Now suppose the relation $(18)$ holds at time $t_{k-1}$, there are two cases at time $t_{k}$, case 1: $N(t_k) = N(t_{k-1})$, and case 2: $N(t_k) > N(t_{k-1})$. First let us consider the case when $N(t_k) = N(t_{k-1})$. In this case, we know that $\Delta_{a}^n(\bar{s}^n(t_{k-1})) = \Delta_{a}^n(\tilde{S}(t_{k-1})) := \Delta_{a}^n$ is independent of $\bar{s}^n(t_{k-1})$ and $\tilde{S}^n(t_{k-1})$. Also, keep in mind that $|\Delta_{a}^n(t_k)| \leq 1/\sqrt{n}$. Now we can write the increment of the difference process

$$(\bar{s}^n(t_k) - \tilde{S}^n(t_k)) - (\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}))$$

$$= \Delta_{a}^n \vee (\lceil -\bar{s}^n(t_k)/(2\delta) \rceil \delta) + \Delta_{b}^n \vee (\lceil -\bar{s}^n(t_k)/(2\delta) \rceil \delta) - \Delta_{a}^n + \Delta_{b}^n \vee (\lceil -\bar{s}^n(t_{k-1}) \rceil \delta). \quad (19)$$

Therefore, if $\bar{s}^n(t_{k-1}) \geq \tilde{S}^n(t_{k-1}) \geq 2/\sqrt{n}$, we have

$$(\bar{s}^n(t_k) - \tilde{S}^n(t_k)) - (\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})) = \Delta_{a}^n + \Delta_{b}^n - (\Delta_{a}^n + \Delta_{b}^n) = 0$$

and as a result $\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}) = \bar{s}^n(t_k) - \tilde{S}^n(t_k)$. If $\bar{s}^n(t_{k-1}) \geq 2/\sqrt{n} \geq \tilde{S}^n(t_{k-1}) \geq 0$, We have

$$(\bar{s}^n(t_k) - \tilde{S}^n(t_k)) - (\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}))$$

$$= \Delta_{a}^n + \Delta_{b}^n - (\Delta_{a}^n + \Delta_{b}^n) \vee (\lceil -\bar{s}^n(t_{k-1}) \rceil \delta) = - (\tilde{S}^n(t_{k-1}) + \Delta_{a}^n + \Delta_{b}^n) \leq 0.$$
Therefore,
\[
\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}) \geq \bar{s}^n(t_k) - \tilde{S}^n(t_k) \geq \frac{2}{\sqrt{n}} - (\bar{s}^n(t_{k-1}) + \Delta^a_n + \Delta^b_n) \geq 0.
\]
Otherwise, we have \(0 \leq \tilde{S}^n(t_{k-1}) \leq \bar{s}^n(t_{k-1}) \leq 2/\sqrt{n}\). In this case, one can check that for any fixed \(\tilde{S}^n(t_k) = \bar{s}\) and \(\bar{s}^n(t_k) = s\), the increment of the difference process (19) reaches its maximum at \(\Delta^a_n = -1/\sqrt{n}\) and \(\Delta^b_n = -\bar{s} + 1/\sqrt{n}\) and its minimum at \(\Delta^a_n = \Delta^b_n = -[s/2\delta]\). Hence,
\[
\bar{s} - s \leq \left(\bar{s}^n(t_k) - \tilde{S}^n(t_k)\right) - \left(\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})\right) \leq 0 \lor \frac{1}{\sqrt{n}} - \frac{s}{2} - \bar{s}.
\]
Plugging in \(\tilde{S}^n(t_k) = \bar{s}\) and \(\bar{s}^n(t_k) = s\), we have
\[
0 \leq \bar{s}^n(t_k) - \tilde{S}^n(t_k) \leq (s - \bar{s}) \lor \left(\frac{1}{\sqrt{n}} + \frac{s}{2}\right) \leq (\bar{s}^n(t_k) - \tilde{S}^n(t_k)) \lor \frac{2}{\sqrt{n}}.
\]
The last inequality holds as \(s = \bar{s}^n(t_{k-1}) \leq 2/\sqrt{n}\). In summary, we have proved that when \(N(t_k) = N(t_{k-1})\), if the relation (18) holds at time \(t_{k-1}\), so does it at time \(t_k\). Now if \(N(t_k) \geq N(t_{k-1}) + 1\), intuitively, at least one jump occurs in \(\Delta^a_n\) and \(\Delta^b_n\). If \(I^a_k = 1\) we have
\[
\Delta^a_n(\tilde{S}^n(t_{k-1})) = I^a_k V^a_k[\tilde{S}^n(t_{k-1})/(2\delta)]\delta, \quad \text{and} \quad \Delta^a_n(\bar{s}^n(t_{k-1})) = I^a_k V^a_k[\bar{s}^n(t_{k-1})/(2\delta)]\delta.
\]
If in addition \(I^b_k = 1\), then
\[
\Delta^b_n(\tilde{S}^n(t_{k-1})) = I^b_k V^b_k[\tilde{S}^n(t_{k-1})/(2\delta)]\delta, \quad \text{and} \quad \Delta^b_n(\bar{s}^n(t_{k-1})) = I^b_k V^b_k[\bar{s}^n(t_{k-1})/(2\delta)]\delta,
\]
and therefore,
\[
\bar{s}^n(t_k) - \tilde{S}^n(t_k) \leq (\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}))(I^a_k V^a_k + I^b_k V^b_k + 1)/2.
\]
As
\[
0 \leq V^a_k + V^b_k + 1 \leq 2c + 1 \quad \text{and} \quad 0 \leq \bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}) \leq ((c + 1)^{N(t_{k-1})} - 1) \cdot \frac{2}{\sqrt{n}},
\]
by the induction assumption we have
\[
0 \leq \bar{s}^n(t_k) - \tilde{S}^n(t_k) \leq ((c + 1)^{N(t_{k-1})} - 1)(2c + 1) \cdot \frac{2}{\sqrt{n}} \leq ((c + 1)^{N(t_k)} - 1) \cdot \frac{2}{\sqrt{n}}.
\]
If \(I^b_k = 0\), then following a similar argument as in the case when \(N(t_k) = N(t_{k-1})\), we have
\[
\bar{s}^n(t_k) - \tilde{S}^n(t_k) \leq (\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1}))(V^a_k + 1) + (\bar{s}^n(t_{k-1}) - \tilde{S}^n(t_{k-1})) \lor \frac{2}{\sqrt{n}}
\]
\[
= ((c + 1)^{N(t_k)} - 1) \cdot \frac{2}{\sqrt{n}} \quad \text{as} \quad c \geq 1.
\]
In summary, we have proved the relation (18) of \(\tilde{S}^n(\cdot)\) and \(\bar{s}^n(\cdot)\) by induction. Now let us turn to the difference \(\tilde{m}^n(\cdot) - \bar{M}^n(\cdot)\). Actually, \(\tilde{m}^n(t) - \bar{M}^n(t)\) can be decomposed into two parts,
\[
\tilde{m}^n(t) - \bar{M}^n(t) \leq \sum_{0 \leq k \leq [nt]: N(t_{k+1}) = N(t_k)} \left[\Delta^a_n(t_k) \lor ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta^b_n(t_k) \lor ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - (\Delta^a_n(t_k) - \Delta^b_n(t_k))\right]
\]
\[
+ \sum_{i=1}^{N(t)} ([\bar{s}^n(t_{k-1})/(2\delta)]\delta - [\tilde{S}^n(t_{k-1})/(2\delta)]\delta)(I^a_k V^a_k + I^b_k V^b_k),
\]
38
where \( \{t_{ki}\} \) are the jump times. We denote the two summation parts as

\[ \bar{m}^n(t) - \bar{M}^n(t) = \epsilon^n_0(t) + \epsilon^n_1(t). \]

Intuitively, \( \epsilon^n_0(t) \) is the error corresponding to the diffusion part when \( I^n_k = I^n_{k+1} = 0 \) and \( \epsilon^n_1(t) \) is the error corresponding to the jumps. In the summation part \( \epsilon^n_0(t) \), we write \( \Delta^n_0(\bar{s}^n(t_k)) = \Delta^n_0(\bar{S}^n(t_k)) \), because they are independent of \( s^n(t_k) \) and \( S^n(t_k) \) when when \( I^n_k = I^n_{k+1} = 0 \). Following a same induction argument as for \( \bar{s}^n - \bar{S}^n \), we can show that the error caused by jumps \( \epsilon^n_1(t) \) satisfies that

\[ \epsilon^n_1(t) \leq ((1 + 2c)^N(t) - 1) \cdot \frac{2}{\sqrt{n}}. \]

On the other hand, note that \( \epsilon^n_0(t) \) equals

\[
\sum_{0 \leq k \leq [nt]: N(t_{k+1}) - N(t_k)} \left[ \Delta_0^n(t_k) \vee ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_0^n(t_k) \vee ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - (\Delta_0^n(t_k) - \Delta_0^n(t_k)) \right] = \sum_{0 \leq k \leq [nt]: N(t_{k+1}) - N(t_k)} \left[ (\Delta_0^n(t_k) \vee ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_0^n(t_k)) - (\Delta_0^n(t_k) \vee ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_0^n(t_k)) \right].
\]

Since \( \Delta_0^n(t_k) \) and \( \Delta_0^n(t_k) \) are independent and identically distributed, we have that for any \( k \geq 1 \)

\[
E[\epsilon^n_0(t_k) - \epsilon^n_0(t_{k-1})|\mathcal{F}^n_{t_k}] = P(N(t_k) = N(t_{k-1})) \cdot (E[\Delta_0^n(t_k) \vee ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_0^n(t_k)] - E[\Delta_0^n(t_k) \vee ([\bar{s}^n(t_{k-1})/(2\delta)]\delta) - \Delta_0^n(t_k)]) = 0,
\]

where \( \mathcal{F}^n_{t_k} \) is the \( \sigma \)-field generated by \( \{\Delta_0^n(t_i), \Delta_0^n(t_i), \bar{S}^n(t_i)\}^k_{i=1} \). Therefore, the process \( \epsilon^n_0(\cdot) \) is a martingale under the filtration \( \mathcal{F}^n \). Besides, as \( |\Delta_0^n(t_k)| \leq 1/\sqrt{n} \) when \( N(t_k) = N(t_{k-1}) \), we have

\[
|\epsilon^n_0(t_k) - \epsilon^n_0(t_{k-1})| \leq \frac{2}{\sqrt{n}} I(\bar{s}^n(t_{k-1}) < \frac{2}{\sqrt{n}}).
\]

The quadratic variation

\[
[\epsilon^n_0](t) \leq \frac{4}{n} \sum_{i=0}^{[nt]} I(\bar{s}^n(t_i) < \frac{2}{\sqrt{n}}).
\]

Recall that we have proved \( \bar{s}^n(\cdot) \geq \bar{S}^n(\cdot) \),

\[
[\epsilon^n_0](t) \leq \frac{4}{n} \sum_{i=0}^{[nt]} I(\bar{S}^n(t_i) < \frac{2}{\sqrt{n}}).
\]

Since \( 2/\sqrt{n} \to 0 \), for any \( \zeta > 0 \) we have

\[
\lim_{n \to \infty} [\epsilon^n_0](t) \leq \lim_{n \to \infty} 4 \int_0^t I(\bar{S}^n(u) < \zeta)du \leq \lim_{n \to \infty} 4 \int_0^t f^\zeta(\bar{S}^n(u))du,
\]

where \( f^\zeta(\cdot) \) is a smooth function on \( \mathbb{R}^+ \) and satisfies \( f(x) = 1 \) for all \( 0 \leq x \leq \zeta, 0 \leq f(x) \leq 1 \) for \( \zeta \leq x \leq 2\zeta \) and \( f(x) = 0 \) for \( x > 2\zeta \). (Such function can be constructed, for instance, by convolution.) Since \( f^\zeta(\cdot) \) is bounded and \( \bar{S}^n(\cdot) \) converges weakly to the limit process (8), we have

\[
\lim_{n \to \infty} E[[\epsilon^n_0](t)] \leq \lim_{n \to \infty} 4E[\int_0^t f^\zeta(\bar{S}^n(u))du] = 4E[\int_0^t f^\zeta(\bar{s}(u))du] \leq 4E[\int_0^t I(\bar{s}(u) \leq 2\zeta)],
\]
As the limit process \( \bar{s}(\cdot) \) has the same dynamics as a reflected Brownian motion except when at the finite time of jumps on \([0, t]\), we have \( E[\int_0^t I(\bar{s}(u) \leq 2\zeta)] \to 0 \) as \( \zeta \to 0 \). Since \( \zeta \) can be arbitrarily small, we conclude that the expected quadratic variation \( E[\epsilon_0^2(t)] \to 0 \) as \( n \to 0 \) for any \( t < \infty \).

By Doob’s Inequality, we have that for all fixed \( \zeta > 0 \),
\[
P(\max_{0 \leq u \leq t} |\epsilon_0^n(u)| > \zeta) \leq \frac{E[\epsilon_0^n(t)]}{\zeta^2} \to 0.
\]

Therefore, \( \epsilon_0^n(\cdot) \) converges weakly to \( x(\cdot) \equiv 0 \) in space \( D[0, t] \) for all \( t < \infty \). In the end, the counting process \( N(\cdot) \) converges to a time inhomogeneous Poisson process with rate \( \gamma \bar{\mu}(t) \), where \( \bar{\mu}(t) \) is given in the proof of Proposition 1. Therefore for any \( t < \infty \)
\[
E[\max_{0 \leq u \leq t} |((2c + 1)^N(u) - 1)\frac{2}{\sqrt{n}}|] = E[((2c + 1)^{N(t)} - 1)\frac{2}{\sqrt{n}}] = O(\frac{1}{\sqrt{n}}).
\]

As a result, the process \( ((2c + 1)^{N(\cdot)} - 1)\frac{2}{\sqrt{n}} \) converges weakly to \( x(\cdot) \equiv 0 \) in space \( D[0, t] \). Recall that we have proved that \( ((2c + 1)^N(\cdot) - 1)\frac{2}{\sqrt{n}} \) is an upper bound of \( |\bar{s}^n(\cdot) - \bar{S}^n(\cdot)| \) and the ‘jump part’ of \( |\bar{m}^n(\cdot) - \bar{M}^n(\cdot)| \). As a consequence, we can conclude that the difference process \( (\bar{s}^n(\cdot) - S^n(\cdot), \bar{m}^n(\cdot) - \bar{M}^n(\cdot)) \) converges weakly to \((0, 0)\) on any compact interval \([0, t]\). \( \blacksquare \)

References


