A MARKOV MODEL OF A LIMIT ORDER BOOK: THRESHOLDS, RECURRENCE, AND TRADING STRATEGIES

FRANK KELLY AND ELENA YUDOVINA

Abstract. We formulate an analytically tractable model of a limit order book on short time scales, where the dynamics are driven by stochastic fluctuations between supply and demand and order cancellation is not a prominent feature. We establish the existence of a limiting distribution for the highest bid, and for the lowest ask, where the limiting distributions are confined between two thresholds. We make extensive use of fluid limits in order to establish recurrence properties of the model. We use our model to analyze various high-frequency trading strategies, and the Nash equilibrium that emerges between high-frequency traders when a market continuous in time is replaced by frequent batch auctions.

1. Introduction

A limit order book (LOB) is a trading mechanism for a single-commodity market. The mechanism is of significant interest to economists as a model of price formation. It is also used in many financial markets, and has generated extensive research, both empirical and theoretical: for a recent survey, see Gould et al. [10].

The detailed historic data from LOBs in financial markets has encouraged models able to replicate the observed statistical properties of these markets. Unfortunately, the added complexity usually makes the models less analytically tractable. In particular, with relatively few exceptions, models of limit order books are only amenable to simulation or numerical exploration. Our aim in this paper is to develop a simple and analytically tractable model of a LOB. We do this by explicitly excluding from the model a number of significant features of real-world markets. We shall see that, even in our simple model, a number of non-trivial and insightful results can be obtained, with implications for more complex models.

We next describe an example of our model of a LOB, and our results for the example. A *bid* is an order to buy one unit, and an *ask* is an order to sell one unit. Each order has associated with it a *price*, a real number. Suppose that bids and asks arrive as independent Poisson processes of unit rate and that the prices associated with bids, respectively asks, are independent identically distributed random variables with density $f_b(x)$, respectively $f_a(x)$. An arriving bid is either added to the LOB, if it is lower than any asks present in the LOB, or it is matched to the lowest ask and both depart. Similarly an arriving ask is either added to the LOB, if it is higher than any bids present in the LOB, or it is matched to the highest bid and both depart. The LOB at time t is thus the set of bids and asks (with their prices), and our assumptions imply the LOB is a Markov process.

For this model we show that there exists a threshold κ_b with the following properties: for any $x < \kappa_b$ there is a finite time after which no arriving bids less than x are ever matched; and for any $x > \kappa_b$ the event that there are no bids greater than x in the LOB is recurrent. Similarly, with directions of inequality reversed, there exists a corresponding threshold κ_a for asks. Further there is a density $\pi_a(x)$, respectively $\pi_b(x)$, supported on (κ_b, κ_a) giving the limiting distribution of the lowest ask, respectively highest bid, in the LOB. The densities π_a, π_b solve the equations

$$f_b(x) \int_x^{\kappa_a} \pi_a(y)dy = \pi_b(x) \int_{-\infty}^{x} f_a(y)dy \tag{1a}$$

$$f_a(x) \int_{\kappa_b}^{x} \pi_b(y)dy = \pi_a(x) \int_{x}^{\infty} f_b(y)dy. \tag{1b}$$

Key words and phrases. limit order book, queueing, fluid limit, high-frequency trading .

The second author's research was partially supported by NSF Graduate Research Fellowship and NSF grant DMS-1204311.

For example, if $f_a(x) = f_b(x) = 1, x \in (0,1)$, then $\kappa_a = \kappa, \kappa_b = 1-\kappa$, $\pi_a(x) = \pi_b(1-x)$, and

$$\pi_b(x) = (1-\kappa)\left(\frac{1}{x} + \log\left(\frac{1-x}{x}\right)\right), \quad x \in (\kappa, 1-\kappa) \tag{2}$$

where the value of κ is given as follows. Let w be the unique solution of $we^w = e^{-1}$: then $w \approx 0.278$ and $\kappa = w/(w+1) \approx 0.218$. Observe that any example with $f_a = f_b$ can be reduced to this example by a monotone transformation of the price axis.

The existence of thresholds with the claimed properties is a relatively straightforward result, using Kolmogorov's 0–1 law. In order to make the claimed distributional result precise the major challenge is to establish positive recurrence of certain *binned* models: such models arise naturally where, for example, prices are recorded to only a finite number of decimal places. Given a sufficiently strong notion of recurrence the intuition behind equations (1) is straightforward: in equilibrium the right-hand side of equation (1a) is the probability flux that the highest bid in the LOB is at x and that it is matched by an arriving ask with a price less than x, and the left-hand side is the probability flux that the lowest ask in the LOB is more than x and that an arriving bid enters the LOB at price x; these must balance, and a similar argument for the lowest ask leads to equation (1b). To establish positive recurrence of the binned models we make extensive use of fluid limits [2], an important technique in the study of queueing networks.

The orders we have described so far are called *limit orders* to distinguish them from *market orders* which request to be fulfilled immediately at the best available price. Market orders are straightforward to include in our model: in the example just described we simply associate a price 1 or 0 with a market bid or market ask respectively. As the proportion of market bids increases towards a critical threshold, $w \approx 0.277$ in the above example, the support of the limiting distributions π_a, π_b increases to approach the entire interval $(0,1)$: above the threshold the model predicts recurring periods of time when there will exist either no highest bid or no lowest ask in the LOB. This conclusion holds, with the same critical threshold w, for any example with $f_a = f_b$.

Our model of a LOB is a form of two-sided queue, the study of which dates at least to the early paper of Kendall [11] who modelled a taxi-stand with arrivals of both taxis and travellers as a symmetric random walk. Recent theoretical advances involve servers and customers with varying types and constraints on feasible matchings between servers and customers, with applications ranging from large-scale call centres to national waiting lists for organ transplants [1, 15, 16]. Our interest in models of LOBs is in part due to the simplicity of the matchings in this particular application: types, as real variables, are totally ordered and so when an arriving order can be matched the match is uniquely defined.

Next we comment on several important features of real-world markets that are missing from our basic model of a LOB. We assume that orders are never cancelled and that the arrival streams of orders, with their prices, are not dependent on the state of the LOB. These assumptions might correspond with orders from long-term investors who place orders for reasons exogenous to the model,[1] and who view the market as effectively efficient for their purposes. These assumptions, and the related assumption of stationarity of the arrival streams, may be more reasonable for a high-volume market where there may be a substantial amount of trading activity even over time periods where no new information becomes available concerning the fundamentals of the underlying asset. Mathematically our model may be viewed as assuming a separation between these time-scales. We also assume that all orders are for a single unit. Here we note that an investor who is attempting to be passive in her execution, so as not to move the price against her, may well spread a larger order in line with volume in the market [7].

Markets may contain traders other than long-term investors, and there is currently considerable interest in the effect of high-frequency trading on LOBs. Interestingly, many high-frequency trading strategies are straightforward to represent within our model, since traders who can react immediately to an order entering the LOB may leave the Markov structure intact. Consider first the following *sniping* strategy for a single high-frequency trader: he immediately buys every bid that joins the LOB at price above q and every ask that joins the LOB at price below p, where p and q are chosen to balance the rates of these purchases. This model fits straightforwardly within our framework, and we show how to calculate the optimal values, for the high-frequency trader, of the constants p and q. A single trader might instead behave as a *market maker*

[1] For example, to manage their portfolios. Investors may differ in their preferences and in their valuations, even given the same information, which creates potential gains from trade.

and place an infinite number of bid, respectively ask, orders at p, respectively q, where $\kappa_b < p < q < \kappa_a$. We are again able to analyze this case. Notably the optimal rate of return under this strategy may beat that under the sniping strategy: it does so for the simple example above where $f_a(x) = f_b(x) = 1, x \in (0, 1)$, describes the order flow from long-term investors. But a third strategy, which combines market making and sniping, will generally beat both the individual strategies.

Our model also allows us to study the Nash equilibrium when there are multiple high-frequency traders competing using mixtures of market making and sniping. There has been considerable discussion recently of the effects of competition between multiple high-frequency traders, and of proposals aimed to slow down markets. A key issue is that high-frequency traders may compete on the speed with which they can snipe an order, rather than compete on price. Budish et al. [3, 4] propose replacing a market continuous in time with frequent batch auctions, held perhaps several times a second. We study the Nash equilibrium in such a market when there are multiple high-frequency traders competing using mixtures of market making and sniping. Competition removes the attraction for the traders of market making, and the Nash equilibrium has traders sniping bids above, respectively asks below, a central price.

The previous research most similar in mathematical framework to that reported here is by Cont and coauthors [6, 5] who also describe LOBs as a Markovian system of interacting queues, and are able to obtain analytical expressions for various quantities of interest. In their model the arrival rates of orders at any given price depends on how far the price is above or below the current best ask or bid price. Gao et al. [9] study the temporal evolution of the the shape of a LOB in the model of [6], under a scaling limit. The work of Lachappelle et al. [12], building on Roşu [14], uses a different mathematical framework, that of a mean field game, but shares with our approach some important features. In particular, these authors distinguish between institutional investors whose decisions are independent of the immediate state of the LOB and high-frequency traders who trade as a consequence of the immediate state of the LOB. The models of both [5] and [12] keep detailed information on queue sizes only at the best bid and best ask prices; [6] shares with our approach a Markov description of the entire LOB.

In much of the market microstructure literature features of LOBs, such as large bid-ask spreads, are explained as a consequence of participants protecting themselves from others with superior information. While this is clearly an important aspect of real-world markets we note that such features may also arise for other reasons. The driving force for the dynamics of the LOB in our approach, as in [12, 14], is not asymmetric information but stochastic fluctuations between supply and demand.

The organisation of the paper is as follows. In Section 2 we describe precisely our model and our main results. Section 3 develops the scaffolding necessary for the proofs, which are given in Section 4. In Section 5 we describe some applications of our results: this section contains our discussion of market orders and of high-frequency trading strategies.

2. Model and results

Informally, the limit order book operates as follows. There are some queued orders, called "bids" and "asks"; a bid is an order to buy a unit of a commodity for at most a given price, and an ask is an order to sell a unit of commodity for at least a given price. New orders arrive into the book, and whenever there is a price match, orders are executed: the newly arriving order departs with *the best order* on the opposite side of the book. We now formalize this notion.

The state of the LOB at time t is a pair (B_t, A_t) of (possibly infinite) counting measures on $\mathbb{R}$. The counting measure B_t represents the prices of queued bids, while A_t represents the prices of queued asks. New orders arrived as a labeled Poisson point process; [2] each label records the type of order (bid or ask) and the price (any real number). We assume the labels of orders are independent and identically distributed, and in particular independent of the state of the book. We allow the price distributions of bids and asks to be different.

We formalize the concept of a "price match" using a *price function*, that is a nondecreasing, not necessarily continuous, function $\mathcal{P} : \mathbb{R} \to \mathbb{R}$. A bid-ask pair is *weakly compatible* if $\mathcal{P}(\text{bid}) \geq \mathcal{P}(\text{ask})$, and *strongly compatible* if $\mathcal{P}(\text{bid}) > \mathcal{P}(\text{ask})$. The particularly important examples of $\mathcal{P}$ will be $\mathcal{P}(x) = x$, and the

[2] It will not be important that the time structure is that of a Poisson process, only that there is a sequence of incoming orders.

function that partitions all prices into n pricing bins; we will refer to this latter case, where the image of $\mathcal{P}$ is a finite set, as the *binned model*. Note that compatibility of bid-ask pairs is unchanged under any strictly increasing transformation of prices; so without loss of generality we may assume that $\mathcal{P}(x) \in (0,1)$ for all x. We will therefore think of prices as falling in the interval $(0,1)$.

We are now ready to define the *strict* and *non-strict* limit order books $\overline{\mathcal{L}}_t$ and $\mathcal{L}_t$.

Initial state: Initially, there should be no compatible bid–ask pairs in the book. Equivalently, the initial state (B_0, A_0) satisfies

$$B_0(x,1) \cdot A_0(0,y) = 0 \quad \text{if} \quad \begin{cases} \mathcal{P}(y) < \mathcal{P}(x), & \text{strict} \\ \mathcal{P}(y) \le \mathcal{P}(x), & \text{non-strict} \end{cases}$$

Change at order arrival: We do not allow cancellations in the model, so all changes to the state occur at the time of an order arrival. Suppose at time t a bid at price p arrives. If there is a matching ask in the book, i.e. if $A_{t-}(0,y) > 0$ for some y such that $\mathcal{P}(y) < \mathcal{P}(p)$ (strict) or $\mathcal{P}(y) \le \mathcal{P}(x)$ (non-strict), then nothing happens to the bids in the book ($B_t = B_{t-}$), and the lowest ask departs: $A_t = A_{t-} - \delta_q$, where $q = \min\{x : A_{t-}\{x\} > 0\}$. If there are no matching asks in the book, the bid joins the book: $B_t = B_{t-} + \delta_p$ and $A_t = A_{t-}$. The situation is symmetric if the arriving order is an ask at price q: if there is a matching bid, the two orders depart (so $A_t = A_{t-}$ and $B_t = B_{t-} - \delta_p$ where $p = \max\{x : B_{t-}\{x\} > 0\}$), and if there are no matching bids, then the ask joins the book ($B_t = B_{t-}$ and $A_t = A_{t-} + \delta_q$).

We will be keeping track of the highest (price of a) bid β_t and lowest (price of an) ask α_t in the book at time t. If an order departs the book at time t, it must be at price β_{t-} (if a bid) or α_{t-} (if an ask). We allow $B_0\{x\} = \infty$ or $A_0\{y\} = \infty$; if this is the case, then no bids left of x, and no asks right of y, will ever depart the limit order book, since they will never be the highest bid (respectively lowest ask).

Below, we will refer to continuous and discretized models of LOBs. A continuous LOB is one where the bid and ask price distributions are absolutely continuous, and the price function is strictly increasing (e.g. $\mathcal{P}(x) = x$). Discretized models will typically have the binned price function. In a continuous LOB, strict and non-strict compatibility coincide, because almost surely all order prices are distinct.

For a discretized, binned LOB, we will use notation $[\![x]\!]$ to denote the index of the bin containing x; $[\![x]\!]$ is a positive integer ranging from 1 to N for some $N > 0$. We may also write $\mathcal{P}([\![x]\!])$ to mean $\mathcal{P}(x)$; there will be no confusion, because $x \in (0,1)$ and $[\![x]\!]$ is an integer.

We now present the main results concerning the model. The first result, Theorem 2.1, establishes a transition at threshold values κ_b and κ_a. Eventually bids arriving below κ_b, and asks arriving above κ_a, will never be executed; whereas all bids arriving above κ_b, and all asks arriving below κ_a, will be executed. The second result, Theorem 2.2, presents the distribution of the rightmost bid and leftmost ask.

Theorem 2.1 (Thresholds)**.** *There exist prices κ_b and κ_a with the following properties:*

1. *Almost surely there exists a random time $T_0 < \infty$ such that $\beta_t > \kappa_b$ and $\alpha_t < \kappa_a$ for all $t \ge T_0$.*
2. *For any $\epsilon > 0$, infinitely often there will be no orders with prices in $(\mathcal{P}(\kappa_b) + \epsilon, \mathcal{P}(\kappa_a) - \epsilon)$.*
3. *Let (x,y) be such that $\mathcal{P}(x) > \mathcal{P}(\kappa_b) + \epsilon$ and $\mathcal{P}(y) < \mathcal{P}(\kappa_a) - \epsilon$ for some $\epsilon > 0$. Consider the LOB started with infinitely many bids at x, infinitely many asks at y, and finitely many orders in between. The evolution of the orders at prices in the interval (x,y) is a positive (Harris) recurrent Markov process, with finite expected time until there are no orders in the interval.*

The existence of a threshold κ_b such that bids leave to the right of it and do not leave to the left of it, at least eventually, is an easy corollary of Kolmogorov's 0–1 law. The challenging part of the above theorem is the claim that in certain intervals there will *simultaneously* be neither bids nor asks. In fact, we will need to prove this part of Theorem 2.1 and Theorem 2.2 below simultaneously, and will make extensive use of fluid models.

Theorem 2.2 (Distribution of the highest bid)**.** *Suppose $\mathcal{P}(x) = x$ and suppose the arrival price pdfs f_b and f_a are supported on $[0,1]$, are bounded above and below, are absolutely continuous, and have bounded densities bounded above and below: $M^{-1} \le f^a, f^b \le M$. Let π^b be the limiting density of the highest bid, and let $\varpi^b = \pi^b / f^b$; define π^a and ϖ^a similarly. Then $\kappa_b > 0$, $F^b(\kappa_b) = 1 - F^a(\kappa_a)$, and ϖ^b satisfies the*

following ODE:

$$\left(-\frac{f_a(x)}{1-F^b(x)}(F^a(x)\varpi^b(x))'\right)' = \varpi^b(x)f_b(x)$$

with initial conditions

$$(F^a(x)\varpi^b(x))|_{x=\kappa_b} = 1, \qquad (F^a(x)\varpi^b(x))'|_{x=\kappa_b} = 0$$

and the additional constraint $\varpi^b(x) \to 0$ as $x \uparrow \kappa_a$. The distribution of the lowest ask satisfies a similar ODE.

Corollary 2.3 (Uniform arrivals)**.** *Suppose that $\mathcal{P}(x) = x$ and the arrival price distribution is uniform on $[0,1]$ for both bids and asks. Then $\kappa_b \approx 0.217$ is given by $\kappa = w/(w+1)$ where $we^w = e^{-1}$. The limiting density of the highest bid is supported on $(\kappa_b, 1-\kappa_b)$, and is given by*

$$\varpi^b(x) = \mathbf{1}_{(\kappa,1-\kappa)}(1-\kappa)\left(\frac{1}{x} + \log\left(\frac{1-x}{x}\right)\right)$$

and the limiting density of the lowest ask is $\varpi^a(x) = \varpi^b(1-x)$.

Remark 1 (Absolute continuity). The assumption of absolute continuity of F^b and F^a with respect to the Lebesgue measure can be replaced by the assumption that the derivative dF^b/dF^a exists and is bounded above and below. Indeed, whenever this is the case, we can replace F^b and F^a by absolutely continuous distributions, and use a suitable price function to transform them. (However, the result of Theorem 2.2 is easier to state in terms of densities.) The requirement that the ratio of the densities be bounded avoids the trivial examples $f^b = 2\mathbf{1}_{[0,1/2)}$, $f^a = 2\mathbf{1}_{(1/2,1]}$ (nonoverlapping supports, no orders leave) or $f^a = 2\mathbf{1}_{[0,1/2)}$, $f^b = 2\mathbf{1}_{(1/2,1]}$ (nonoverlapping supports, no threshold). It may be possible to relax the requirement on bounded derivatives and simply require the supports of the distributions to coincide.

As long as the arrival price distributions are absolutely continuous as above we can reparametrize space so that bids arrive uniformly on the interval $[0,1]$, and so Corollary 2.3 covers also the case where the distributions of arriving bid and ask prices are identical. We describe some other analytically tractable examples of Theorem 2.2 in Section 5.

In Figure 1, we show the exact limiting distribution of the highest bid for the non-strict binned LOB with uniform arrivals 50 bins, along with the limiting distribution for the continuous LOB. Note the "shoulder" bin: in the binned LOB, the threshold happens to fall into the middle of a bin, so the long-term probability of having the rightmost bid land into the bin is positive but below the limiting prediction.

While we have been able to compute analytically the distribution of the location of the rightmost bid, there are many related quantities for which we do not have an exact expression (although the positive recurrence established in Theorem 2.1 implies that they are well-defined and can be estimated consistently from simulation). Notably, we have not been able to derive analytic expressions for the equilibrium height of the book (i.e. expected number of bids or asks at a given price in the binned model), or for the joint distribution of the highest bid and lowest ask.

3. Preliminary results: monotonicity

Before proving the main results, we erect a certain amount of scaffolding. Part of its purpose is to allow us to transition between continuous LOBs (for which we expect to get differential equations in the answer) and binned models (which can be modelled as countable-state Markov chains). It also allows us to compare LOBs with different arrival price distributions.

Lemma 3.1 asserts that the state of the limit order book is Lipschitz in the initial state with Lipschitz constant 1: in particular, small perturbations in the arrival and matching patterns will lead to small perturbations in the state of the book. Lemma 3.2 asserts that actions that decrease cumulative bid and ask queues by either shifting orders or removing them in bid–ask pairs will only decrease future queue sizes.

Lemma 3.1 (Adding one order)**.** *Consider a limit order book $\mathcal{L}$, and let $\tilde{\mathcal{L}}$ differ from $\mathcal{L}$ by the addition of one bid at time 0; let their arrival processes and matching function be the same. Then at all times $\tilde{\mathcal{L}}$ differs from $\mathcal{L}$ either by the addition of one bid, or by the removal of one ask.*

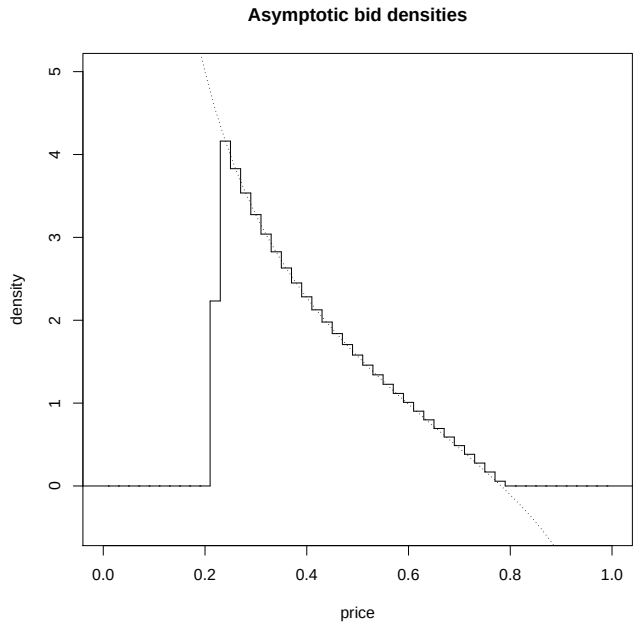

FIGURE 1. Limiting density of the highest bid for the non-strict binned LOB with 50 bins, and limiting density for a continuous LOB (dotted line). Note the "shoulder" bin containing the threshold in the binned model.

Proof. The roles of "bid" and "ask" are symmetric here. The claim clearly holds until the additional bid is the highest bid that departs from the system; once it does, $\mathcal{L}$ differs from $\tilde{\mathcal{L}}$ by the addition of a single ask, and the result follows by induction. □

Define cumulative queue sizes $Q^b(p,t) = B_t(0,p]$, $Q^a(p,t) = A_t[p,1)$. (Note that we count bids from the left and asks from the right.) When we want to highlight the dependence on only one of the variables, we will drop the other variable into a subscript.

Lemma 3.2 (Decreasing queues)**.** *Consider a limit order book $\mathcal{L}$, and let $\tilde{\mathcal{L}}$ differ from $\mathcal{L}$ by modifying the initial state in such a way that $\tilde{Q}^b_0 \le Q^b_0$, $\tilde{Q}^a_0 \le Q^a_0$ (as functions of p), and also $\tilde{Q}^b(1) - \tilde{Q}^a_0(0) = Q^b_0(1) - Q^a_0(0)$. In words, to get from $\mathcal{L}$ to $\tilde{\mathcal{L}}$, at time 0 we remove some bid–ask pairs, and/or shift some bids to the right, and/or shift some asks to the left. Then at all future times $t \ge 0$, $\tilde{Q}^b_t \le Q^b_t$ and $\tilde{Q}^a_t \le Q^a_t$.*

Proof. We show $\tilde{Q}^b \le Q^b$, the argument for asks being identical. The argument proceeds by induction on time, i.e. the number of arrived orders.

Consider first the arrival of a bid at time $t+$ and price p. For it to upset the inequality, it must stay in $\tilde{\mathcal{L}}$ but depart immediately in $\mathcal{L}$; additionally, we need $Q^b_t(q) = \tilde{Q}^b_t(q)$ for some $q \ge p$. Note that if the bid departs immediately in $\mathcal{L}$, the leftmost ask at α_t must be compatible with p, and in particular there are no bids right of p: $Q^b_t(p) = Q^b_t(\infty)$. This, together with $Q^b_t(q) = \tilde{Q}^b_t(q)$ and $\tilde{Q}^b_t \le Q^b_t$, implies that $Q^b_t(\infty) = \tilde{Q}^b_t(\infty)$. Since bid–ask departures occur in pairs, this in turn implies $Q^a_t(-\infty) = \tilde{Q}^a_t(-\infty)$. But it is easy to see that if $\tilde{Q}^a_t \le Q^a_t$ and they are equal at $-\infty$, then $\tilde{\alpha}_t$ (the leftmost jump of $\tilde{Q}^a_t$) and α_t (the leftmost jump of Q^a_t) satisfy $\tilde{\alpha}_t \le \alpha_t$, and hence the arriving bid actually departs immediately in $\tilde{\mathcal{L}}$ as well.

Next consider the arrival of an ask at time $t+$ and price p. For it to upset the inequality, it must cause the departure of the highest bid in $\mathcal{L}$, but not in $\tilde{\mathcal{L}}$, and we must have $Q^b_t(q) = \tilde{Q}^b_t(q)$ for some $q \ge \beta_t$ with $\mathcal{P}(\beta_t) \ge \mathcal{P}(p)$. Now, in $\tilde{\mathcal{L}}$ there are no bids at prices $\ge \mathcal{P}(p)$, hence $\tilde{Q}^b_t(\infty) = \tilde{Q}^b_t(q) = \lim_{\epsilon\to 0} \tilde{Q}^b_t(\beta_t - \epsilon)$. However, this contradicts the inequality $\tilde{Q}^b_t \le Q^b_t$, since $\lim_{\epsilon\to 0} Q^b_t(\beta_t - \epsilon) = Q^b_t(q) - 1$. □

We now have a way to bound continuous LOBs by binned models as follows. Notice that in a non-strict LOB, merging bins results in extra order pairs departing from the LOB, which decreases queue sizes; in a strict LOB, this happens if we split bins. A continuous LOB is simultaneously strict and non-strict; so we

can bound the queues in a continuous LOB by a strict binned LOB from above, and by a non-strict binned LOB from below. We can also bound the order queues of a continuous LOB from above by a book in which all bid arrivals are shifted slightly to the left, and all ask arrivals are shifted slightly to the right. Notice that we can reproduce the behavior of a strict binned LOB from a non-strict one by shifting all bids one bin left relative to all asks (both initially and as they arrive). Consequently, we can bound a continuous LOB from both sides by non-strict binned LOBs with slightly different arrival patterns (and different price functions, since one of these bounding LOBs has more bins than the other). Lemma 3.1 assures that the states of these two non-strict LOBs will be close to each other, giving us tight control over the behavior of the continuous LOB in terms of systems described by countable-state Markov chains.

4. Proof of main results

We begin by stating a weaker form of Theorem 2.1.

Proposition 4.1 (Weak thresholds)**.** *There exist prices κ_b and κ_a with the following properties:*

1. *Almost surely there exists a random time $T_0 < \infty$ such that $\beta_t > \kappa_b$ and $\alpha_t < \kappa_a$ for all $t \geq T_0$.*
2. *For any $\epsilon > 0$, infinitely often there will be no bids with price exceeding $\mathcal{P}(\kappa_b)+\epsilon$. Similarly, infinitely often there will be no asks with price below $\mathcal{P}(\kappa_a) - \epsilon$.*
3. *The threshold values κ_b and κ_a satisfy $F^b(\kappa_b) = 1 - F^a(\kappa_a)$.*

In addition, suppose that the bid and ask price distributions are supported on $[0, 1]$ and are absolutely continuous, with densities bounded above and below. Then the following holds:

4. *For any $\epsilon > 0$, with probability 1, there exists a sequence of times $T_n \to \infty$ such that at time T_n there are no bids with prices above $\mathcal{P}(\kappa_b) + \epsilon$, and the number of asks with prices below $\mathcal{P}(\kappa_a) - \epsilon$ is bounded above by $2M\epsilon T_n$.*

Proof. The first two claims follow from Kolmogorov's 0–1 law. Consider the events

$$\mathcal{E}^b(x) = \{\text{finitely many bids will depart from prices} \leq \mathcal{P}(x)\}$$
$$\mathcal{E}^a(x) = \{\text{finitely many asks will depart from prices} \geq \mathcal{P}(x)\}.$$

Lemma 3.1 shows that these events are in the tail σ-algebra of the arrival process. Since the arrival process consists of a sequence of independent and identically distributed events, Kolmogorov's 0–1 law ensures that for each x, $\mathcal{E}^b(x)$ has probability 0 or 1 (and similarly for $\mathcal{E}^a(x)$). Now let

$$\kappa_b = \sup\{x : \mathbb{P}(\mathcal{E}^b(x)) = 1\}, \qquad \kappa_a = \inf\{x : \mathbb{P}(\mathcal{E}^a(x)) = 1\}. \tag{3a}$$

The first two asserted properties now follow upon noticing that $\mathcal{E}^b(x) \subseteq \mathcal{E}^b(y)$ for $x \geq y$, and that whenever there is a bid departure at price x, there must be no bids at prices higher than $\mathcal{P}(x)$. (The situation is similar for asks.)

We next show that $F^b(\kappa_b) + F^a(\kappa_a) = 1$. From the strong law of large numbers for the arrival process and the 0–1 law above, we know that $F^b(\kappa_b)$ is the proportion of arriving bids that stay in the system:

$$F^b(\kappa_b) = \lim \inf_{t\to\infty} \frac{1}{t}\#(\text{bids in the LOB at time } t). \tag{3b}$$

A similar equality clearly holds for asks with $1-F^a(\kappa_a)$. Since bids and asks always depart in pairs, a further appeal to the strong law of large numbers for the arrival process shows that we must have $F^b(\kappa_b) = 1-F^a(\kappa_a)$.

The existence of times T_n as in part (4) of the theorem follows by a similar argument from the functional law of large numbers for the arrival processes. Picking a large enough time T_n when there are no bids at prices above $\mathcal{P}(\kappa_b) + \epsilon$, we see that there cannot be more than $(1 - F^a(\kappa_a) + M\epsilon)T_n$ asks in the system. Since asks to the right of κ_a arrive at rate $(1-F^a(\kappa_a))$ and eventually never leave, for large enough T_n there will not be more than $2M\epsilon T_n$ asks at prices below $\kappa_a - \epsilon$. □

This result is weaker than the positive recurrence we wish to prove eventually: in particular, it does not show that the total number of orders, both bids and asks, between κ_b and κ_a is ever zero. In order to obtain statements about positive recurrence, we will need to use fluid limit techniques, and our overall approach will be similar to that of [2, Chapter 4]. Specifically, the final proof of stability will come from the use of the multiplicative Foster's criterion (state-dependent drift) [13, Theorem 13.0.1]. In order to get there, we need to show that whenever there are many bids or asks in (κ_b, κ_a), their number decreases at some positive,

bounded below, rate over long periods of time. This is a standard line of argument in queueing theory; but the challenge of our model is that the evolution of the queues depends on which queues are positive, rather than which queues are large. It is known that in general Markov chains of this form are very difficult to analyze ([8] shows that in general the stability of such chains is undecidable), but the special structure of our chain makes it amenable to analysis. The outline of the proof is as follows.

(1) We work with binned LOBs. We begin by showing that, after appropriate rescaling, both the queue sizes and the local time of the highest bid (lowest ask) in each bin converge to a set of Lipschitz trajectories, which we call *fluid limits*. We then proceed to develop properties of the fluid limits.
(2) We next show that all fluid limits tend to zero for bins between $[\![\kappa_b]\!]+1$ and $[\![\kappa_a]\!]-1$. We exploit the equations and inequalities satisfied by fluid limits to show the following:
   (a) There is an interval $[x_0, y_0]$ on which, whenever the fluid limit of the number of orders is positive, it decreases (at a rate bounded below). Therefore, after some time T_0 (which depends on the initial state), the fluid limit will be zero on $[x_0, y_0]$.
   (b) Following T_0, we will be able to bound from below the rate of increase of the local time of the rightmost bid on $[x_1, x_0]$ for some $x_1 < x_0$, and of the leftmost ask on $[y_0, y_1]$ for some $y_1 > y_0$. Since whenever the highest bid is in $[x_1, x_0]$ it has a positive chance of departing, we will conclude that whenever the number of orders in $[x_1, y_1]$ is large, it will decrease (at a rate bounded below). We repeat the argument until $[x_n, y_n] \approx [\kappa_b, \kappa_a]$.
(3) We show that if on some interval, all fluid limits converge to 0 in finite time, then the binned LOB is recurrent on that interval. Since the number of bids in a continuous limit order book can be bounded from above by binned ones, this will also show recurrence of the continuous LOB.

4.1. **ODE of the limiting distribution.** Our first result shows that the ODE which should describe the unique limit, as $t \to \infty$, of the empirical distribution of the highest bid does in fact describe *some* such limit. In the process, we also establish $0 < \kappa_b < \kappa_a < 1$. We consider the case of arrival price pdfs that are bounded from above and below.

Reparametrize space so that $f^a + f^b = 2$ is constant on $[0, 1]$.

Proposition 4.2 (Weak distribution of the highest bid)**.** *Suppose the LOB has N bins, and the arrival price distributions are bounded, absolutely continuous, with densities bounded above and below: $M^{-1} \le f^a, f^b \le M$.*

Let $T_n = T_n(N, \epsilon) \to \infty$ be the sequence of times identified in Proposition 4.1. Let $\pi^b(n, N, \epsilon)$ be the empirical density of the highest bid over the time interval $[0, T_n]$, let π^b be the limit of $\pi^b(n, N, \epsilon)$ as $n, N \to \infty$ and $\epsilon \to 0$. Define also the corresponding densities $\pi^a(n, N, \epsilon) \to \pi^a$ for asks. The limits π^b and π^a exist and are unique. Further, defining $\varpi^b = \pi^b / f^b$ and $\varpi^a = \pi^a / f^a$, these satisfy the pair of integral equations

$$F^a(x)\varpi^b(x) = \int_x^1 \varpi^a(y) f^a(y) dy, \ x \in (\kappa_b, \kappa_a); \qquad \int_{\kappa_b}^{\kappa_a} \varpi^b(x) f^b(x) dx = 1, \tag{4a}$$

$$(1 - F^b(x))\varpi^a(x) = \int_0^x \varpi^b(y) f_b(y) dy, \ x \in (\kappa_b, \kappa_a); \quad \int_{\kappa_b}^{\kappa_a} \varpi^a(x) f^a(x) dx = 1. \tag{4b}$$

Wherever ϖ^b is differentiable, it satisfies the ODE

$$\left(-\frac{1 - F^b(x)}{f_a(x)} (F^a(x)\varpi^b(x))' \right)' = \varpi^b(x) f_b(x) \tag{5a}$$

with initial conditions

$$(F^a(x)\varpi^b(x))|_{x=\kappa_b} = 1, \qquad (F^a(x)\varpi^b(x))'|_{x=\kappa_b} = 0 \tag{5b}$$

and the additional constraint $\varpi^b(x) \to 0$ as $x \uparrow \kappa_a$. The distribution of the leftmost ask satisfies a similar ODE.

Remark 2 (Normalization and initial conditions). (1) From the integral equation (6) it follows that ϖ^b will be continuous, whereas π^b may not be. In particular, if we are interested in piecewise continuous functions f^b and f^a, then ϖ^b will satisfy the ODE on each of the segments where f^b and f^a are

continuous, and can be patched together from the requirement that $\varpi^b(x)$ and $(F^a(x)\varpi^b(x))'$ are both continuous (in x).

(2) The initial conditions here are a consequence of the fact that $\mathbb{P}(\alpha_t < \kappa_b) \to 0$ as $t \to \infty$, which happens for all finite initial states. Consider instead a limit order book with an infinite starting state, e.g. an infinite supply of bids at some price $p > \kappa_b$. Then the initial conditions as above would hold for all $x \in (\tilde{\kappa}_b, p]$, meaning $\varpi^b(x) = 1/F^a(x)$ on that interval. Of course, $\mathbb{P}(\alpha_t \leq p) = 0$. For an LOB with infinite starting state, $\varpi^a(x)$ may not tend to 0 as $x \downarrow p$.

Proof. The proof proceeds as follows:

(1) Fix the number of bins N, and consider the collection of empirical densities $\pi^b(n, N, \epsilon)$, $\pi^a(n, N, \epsilon)$. Along any sequence $n, N \to \infty$ and $\epsilon \to 0$ there is a convergent subsequence.
(2) Any subsequential limit satisfies a certain pair of integral equations, hence some ODEs.
(3) The ODEs will directly imply $\kappa_b < \kappa_a$; in addition, $0 < \kappa_b$ and $\kappa_a < 1$.
(4) The solution to these ODEs is unique, and in particular the limit does not depend on the order of $n, N \to \infty$ and $\epsilon \to 0$.

Step 1:
The space of probability distributions with compact support is compact, so along any sequence of empirical distributions there will be convergent subsequences. Moreover, if the bin width is a, then whenever the highest bid is in bin $[\![x]\!]$, bid departures occur from the bin at rate $\geq F^a(x)\pi^b([\![x]\!])$, whereas bid arrivals occur into that bin at rate at most $f^b(x)$. Consequently, $\pi^b([\![x]\!]) \leq f^b(x)/F^a(x)$ is bounded uniformly in n, N, ϵ, guaranteeing the existence of limiting *densities* along subsequences.

Step 2:
The integral equations are expressing the idea that the rate of bid arrival should be equal to the rate of bid departure. Along a sequence of times where the queues are small (i.e. $\epsilon \approx 0$), this is very nearly true; it will be exactly true in the limit. The bid arrival rate at x is $f^b(x)\mathbb{P}(\alpha_t > x) = f^b(x)\int_x^1 \pi^a(y)dy$, and the bid departure rate at x is $\pi^b(x)F^a(x)$, so setting the two equal gives the result; the ODE is obtained by differentiating twice.

Of course, if we fix N, the limit distribution will be described by a difference equation rather than an integral (or differential) equation. It is standard to see that the limit of solutions to the difference equations solves the differential (or integral) equation.

Step 3:
To see $\kappa_b < \kappa_a$, note that π^b is bounded above by f^b/F^a always, so if it integrates to 1 we must have $\kappa_b < \kappa_a$. To see $\kappa_b > 0$ (and $\kappa_a < 1$), we consider a binned LOB $\tilde{\mathcal{L}}$ with three bins, with bin partitions at x and $x + \epsilon$ for some $x \in (\kappa_b, \kappa_a)$. By monotonicity, $[\![\tilde{\kappa}_b]\!] = 1$ and $[\![\tilde{\kappa}_a]\!] = 3$. For ϵ small enough, the number of orders in the middle bin will eventually be stochastically dominated by a geometric random variable. Indeed, whenever there are bids in bin 2, more bids arrive at rate $F^b(x + \epsilon) - F^b(x)$ and depart at the larger rate $F^a(x + \epsilon)$ (this is after asks from bin 3 stop departing). The situation is similar for asks. Consequently, in $\tilde{\mathcal{L}}$ we must have $\tilde{\pi}^b(2) > 0$ and $\tilde{\pi}^a(2) > 0$. Suppose that $\tilde{\pi}^b(1)$ and $\tilde{\pi}^a(3)$ are such that (almost) all orders depart, then from $\tilde{\pi}^b(1)F^a(x) = F^b(x)$ we find

$$\tilde{\pi}^b(1)F^a(x) = F^b(x) \implies \tilde{\pi}^b(2) = \frac{F^a(x) - F^b(x)}{F^a(x)} \implies F^a(x) > F^b(x).$$

Now let ϵ be small enough that $F^a(x) > F^b(x + \epsilon)$, and solve for $\tilde{\pi}^b(2)$ from the alternative expression $\tilde{\pi}^b(2)F^a(x + \epsilon) = (F^b(x + \epsilon) - F^b(x))\tilde{\pi}^a(3)$. This gives

$$\tilde{\pi}^b(1) + \tilde{\pi}^b(2) = \frac{F^b(x)}{F^a(x)} + \frac{F^b(x + \epsilon) - F^b(x)}{F^a(x + \epsilon)} \frac{1 - F^a(x + \epsilon)}{1 - F^b(x + \epsilon)} < 1.$$

The contradiction shows that in fact in this LOB we must have $\tilde{\pi}^b(1)F^a(x) < F^b(x) - \delta$ for some $\delta > 0$, which implies $F^b(\tilde{\kappa}_b) \geq \delta$. By monotonicity, we obtain $\kappa_b > 0$ as well (for N large enough that the above bin of width ϵ is one of the original bins of the LOB).

Step 4: The uniqueness of solution follows from the fact that we have a second-order ODE with two initial conditions (which, as we just showed, are finite). Note that an alternative argument for $\kappa_b > 0$ would be to show that $\pi^b(x) > 0$ for some $x > 0$, since then the ODE forces $\pi^b F^a / f^b$ decreasing, and $\pi^b(x) \sim 1/x$

near 0, which is not integrable. However, it is not immediately obvious why in a binned LOB the highest bid couldn't spend (almost) all of its time in the leftmost bin, hence we give the more involved argument above. □

The second result we require about the ODE is monotonicity in the initial conditions:

Lemma 4.3 (ODE monotonicity)**.** *Let ϖ^b and $\tilde{\varpi}^b$ be two solutions of the ODE* (5a) *with initial conditions*

$$\varpi^b(x_0) \geq \tilde{\varpi}^b(x_0), \qquad (\varpi^b)'(x_0) \geq (\tilde{\varpi}^b)'(x_0).$$

Then for all $x \geq x_0$, $\varpi^b(x) \geq \tilde{\varpi}^b(x)$.

Proof. Reparametrize space monotonically so that $F^a(x) = x$; this clearly does not affect the result. Then the ODE (5a) becomes

$$(F^b(x) - 1)\left(x(\varpi^b)'(x) + \varpi^b(x)\right)' + x f_b(x)(\varpi^b)'(x) = 0,$$

which is a first-order ODE in $(\varpi^b)'(x)$. Trivially, $\varpi^b(x)$ satisfies a first-order ODE in $(\varpi^b)'(x)$. Since first-order ODEs are increasing in their initial conditions, we obtain the result. □

Corollary 4.4. *Suppose the initial conditions for ϖ^b come from a LOB, and the initial conditions for $\tilde{\varpi}^b$ are $F^a(x_0)\tilde{\varpi}^b(x_0) = 1$, $(F^a(x)\tilde{\varpi}^b(x))'|_{x=x_0} = 0$. Then $\varpi^b(x_0) \leq \tilde{\varpi}^b(x_0)$ and $(\varpi^b)'(x_0) \leq (\tilde{\varpi}^b)'(x_0)$. Consequently, $\varpi^b(x) \leq \tilde{\varpi}^b(x)$.*

Proof. Reparametrize space as before, so that $F^a(x) = x$. Then

$$(x\varpi^b(x))' = -\pi^a(x), \quad \varpi^b(x) = \frac{1}{x}\left(1 - \int_0^x \varpi^a(y)dy\right).$$

From this it is clear that $\varpi^b(x_0) \leq \tilde{\varpi}^b(x_0)$. Further,

$$x(\varpi^b)'(x) = -\pi^a(x) - \varpi^b(x) = -\frac{1}{x} - \int_0^x (\varpi^a(y) - \varpi^a(x))dy.$$

Now, in a LOB, $(1 - F_b(x))\varpi^a(x)$ is increasing (cf. $x\varpi^b(x)$ which is decreasing), meaning ϖ^a is increasing. Consequently, the integral above is nonpositive, and we see $x_0(\varpi^b)'(x_0) \leq -\frac{1}{x_0} = x_0(\tilde{\varpi}^b)'(x_0)$ as required. □

4.2. **Fluid limits.** In this section we introduce the fluid-scaled processes associated with the limit order book, discuss their convergence to fluid limits, and determine properties of the limits.

Throughout the section, we work with a binned limit order book. We assume that bid and ask arrival price distributions are supported on $[0, 1]$ and are absolutely continuous; arriving orders are equally likely to be bids and asks. We may, without loss of generality, reparametrize space and time so that the overall rate of order arrival is the same, $1/N$, for each of the N bins.

Let $B_k(\cdot)$ and $A_k(\cdot)$ be the arrival processes of bids and asks into bin k (indexed by time). For this section, we will assume that arrivals occur at deterministic points of time (one arrival every half of a time unit), and are then assigned the tag (type, price) in an iid fashion. We further assume that the total arrival rate of all orders is 2, so that p_k^b and p_k^a are arrival rates of bids and asks into bin k. Let $Q_k^b(t)$ and $Q_k^a(t)$ be the number of bids, respectively asks, in bin k at time t. Let $T_k^\beta(t)$ and $T_k^\alpha(t)$ be the amount of time up to time t when the rightmost bid, respectively leftmost ask, is in bin k: that is,

$$T_k^\beta(t) = \int_0^t \mathbf{1}\{[\![\beta_s]\!] = k\}ds, \qquad T_k^\alpha(t) = \int_0^t \mathbf{1}\{[\![\alpha_s]\!] = k\}ds.$$

It is clear that the initial data $Q_k^b(0)$, $Q_k^a(0)$ together with the arrival processes $B_k(\cdot)$, $A_k(\cdot)$ give sufficient information to determine the values of all of these processes at later times. In fact, we have the following

expressions:

$$[\![\beta_t]\!] = k \iff Q^b_k(t) > 0, \quad \sum_{k'>k} Q^b_{k'}(t) = 0 \tag{6a}$$

$$[\![\alpha_t]\!] = k \iff Q^a_k(t) > 0, \quad \sum_{k'<k} Q^a_{k'}(t) = 0 \tag{6b}$$

$$Q^b_k(t) = Q^b_k(0) + \int_0^t \mathbf{1}\{[\alpha(s)] > k\} dB_k(s) - \sum_{k' \le k} \int_0^t \mathbf{1}\{[\beta(s)] = k\} dA_{k'}(s) \tag{6c}$$

$$Q^a_k(t) = Q^a_k(0) + \int_0^t \mathbf{1}\{[\beta(s)] < k\} dA_k(s) - \sum_{k' \ge k} \int_0^t \mathbf{1}\{[\alpha(s)] = k\} dB_{k'}(s) \tag{6d}$$

$$T^\beta_k(t) = T^\beta_k(0) + \int_0^t \mathbf{1}\{[\beta(s)] = k\} ds \tag{6e}$$

$$T^\alpha_k(t) = T^\alpha_k(0) + \int_0^t \mathbf{1}\{[\alpha(s)] = k\} ds \tag{6f}$$

We define the *fluid-scaled* processes by $\overline{X}_n(t) = n^{-1}X(nt)$ for any process X.

Let p^b_k and p^a_k be the probabilities that an arriving order is a bid, respectively ask, falling into bin k. (We have $\sum_k (p^a_k + p^b_k) = 1$.) We now have the following result on convergence to fluid limits:

Theorem 4.5 (Convergence to fluid limits)**.** *Consider a sequence of processes*

$$(\overline{B}_{k,n}(\cdot), \overline{A}_{k,n}(\cdot), \overline{Q}^b_{k,n}(\cdot), \overline{Q}^a_{k,n}(\cdot), \overline{T}^\beta_{k,n}(\cdot), \overline{T}^\alpha_{k,n}(\cdot))$$

whose initial state (at time 0) is bounded: $\left\| \overline{Q}^a_{k,n}(0), \overline{Q}^b_{k,n}(0) \right\| \le 1$. *As $n \to \infty$, any such sequence has a subsequence which converges, uniformly on compact sets of t, to a collection of Lipschitz functions*

$$(b_k(\cdot), a_k(\cdot), q^b_k(\cdot), q^a_k(\cdot), t^\beta_k(\cdot), t^\alpha_k(\cdot))$$

uniformly on compact sets. (Different subsequences may converge to different 6-tuples of Lipschitz functions.) We call the limiting 6-tuple a fluid limit.

Any fluid limit satisfies the following equations almost everywhere (i.e. everywhere where the derivatives are defined):

$$b'_k(t) = p^b_k, \quad a'_k(t) = p^a_k \tag{7a}$$

$$t^{\beta\prime}_k(t) = 0 \text{ if } \sum_{k'>k} q^b_{k'}(t) > 0, \quad t^{\alpha\prime}_k(t) = 0 \text{ if } \sum_{k'<k} q^a_{k'}(t) > 0 \tag{7b}$$

$$\sum_{k \le [\![\kappa_b]\!]-1} t^\beta_k(t) = 1, \qquad \sum_{k \ge [\![\kappa_a]\!]+1} t^\alpha_k(t) = 1 \tag{7c}$$

$$q^b_k(t) \ge 0, \quad q^a_k(t) \ge 0 \tag{7d}$$

$$q^{b\prime}_k(t) = 0 \text{ if } q^b_k(t) = 0, \quad q^{a\prime}_k(t) = 0 \text{ if } q^a_k(t) = 0 \tag{7e}$$

$$q^{b\prime}_k(t) = p^b_k \sum_{k'>k} t^{\alpha\prime}_{k'}(t) - t^{\beta\prime}_k(t) \sum_{k' \le k} p^a_{k'} \tag{7f}$$

$$q^{a\prime}_k(t) = p^a_k \sum_{k'<k} t^{\beta\prime}_{k'}(t) - t^{\alpha\prime}_k(t) \sum_{k' \ge k} p^b_{k'}. \tag{7g}$$

Proof. The expression in (6) together with the functional law of large numbers for the arrival processes leads to the u.o.c. convergence along subsequences to a fluid limit. The integral representation implies that limits must be Lipschitz functions.

To see that any fluid limit must satisfy (7), we note that (7a) follows directly for the functional law of large numbers for the arrival processes. Identities (7b) follows from the corresponding statement for prelimit processes: if $\sum_{k'>k} q^b_{k'}(s) > \epsilon > 0$ on a time interval $s \in (t-\epsilon, t+\epsilon)$, then for all sufficiently large n, $\sum_{k'>k} Q^b_{k',n}(ns) > n\epsilon/2 > 0$, so $[\beta(ns)] > k$ and $T^\beta_k(ns)$ is not increasing. Identify (7c) holds for a similar

reason: the rightmost bid (leftmost ask) is always in one of the bins in the prelimit processes, so this must be true in the limit as well. Note that the rightmost bid (leftmost ask) cannot be in any bin $k \geq [\![\kappa_a]\!] - 1$ ($k \leq [\![\kappa_b]\!] + 1$) since there are infinitely many asks (bids) in bin $[\![\kappa_a]\!] - 1$ ($[\![\kappa_b]\!] + 1$). Identities (7d) follows for a similar reason: prelimit queues are nonnegative, hence the limit is nonnegative as well.

Identity (7e) is a corollary of (7d): a process that is always nonnegative, differentiable at t, and equal to 0 at t must have derivative 0 there.

Finally, identities (7f) and (7g) are a corollary of corresponding statements (6) for the prelimit queues, where we make use of (6a) and (6b) in (6). More precisely, the rate at which the bid queue changes is this: if the lowest ask is higher than bin k, then bids arrive into the queue at rate p^b_k; and if the highest bid is in bin k, then all asks arriving at prices below k deplete the queue at k. Because the location of the highest bid or lowest ask does not show up in the fluid limit, we instead use the "local times" t^β and t^α. □

We introduce notation $\pi^\beta_k(t) = (t^\beta_k)'(t)$, $\pi^\alpha_k(t) = (t^\alpha_k)'(t)$.

4.3. **Fluid limits drain.** We will now show that the fluid limit processes drain, that is, converge to 0 on the bins ranging from $[\![\kappa_b]\!] + 1$ to $[\![\kappa_a]\!] - 1$. We will assume that bin widths (and hence p^b_k, p^a_k) are all small.

Theorem 4.6 (Fluid limits drain)**.** *Consider a fluid limit corresponding to a binned LOB with N bins, normalized so that the total arrival rate is 2 (rate 1 for each of the bids and asks). Suppose the arrival process is symmetric ($p^b_k = p^a_{N-k}$), satisfies the absolute continuity requirements, and $p^b(k)$ is decreasing in k (but bounded below, from the absolute continuity requirements). Suppose that initially there are infinitely many bids in bin $[\![\kappa_b]\!]+1$ and infinitely many asks in $[\![\kappa_a]\!]-1$; then the fluid limit of queues can be described by $q^{a,b}_k(t)$ for $[\![\kappa_b]\!] + 2 \leq k \leq [\![\kappa_a]\!] - 2$, and the fluid limit of the local times can be described by $\pi^{a,b}_k(t)$ for $[\![\kappa_b]\!] + 1 \leq k \leq [\![\kappa_a]\!] - 1$.*

Let the initial state of the fluid limit satisfy $\left\|(q^b(0), q^a(0)\right\| \leq 1$. There exists $\epsilon = \epsilon(N) \to 0$ as $N \to \infty$, and a time T depending on $\{M, p^a_k, p^b_k, \text{bin widths}\}$, such that for all bins k satisfying $[\![\kappa_b + \epsilon]\!] < k < [\![\kappa_a - \epsilon]\!]$, and all times $t \geq T$,

$$q^b_k(t) = 0, \qquad q^a_k(t) = 0, \qquad \forall t \geq T.$$

Further, in the interval $[\![\kappa_b+\epsilon]\!] < k < [\![\kappa_a-\epsilon]\!]$ and for $t \geq T$, the derivatives $\pi^\beta_k(t)$ satisfy the second-order difference equation

$$\Delta_k \left(\frac{1 - F^b(k)}{p^a_{k+1}} \cdot \Delta_k \left(\frac{F^a(k)}{p^b_k} \pi^\beta_k \right) \right) = \pi^\beta_{k+1},$$

with initial conditions satisfying

$$\frac{F^a([\![\kappa_b + \epsilon]\!])}{p^b_{[\![\kappa_b+\epsilon]\!]}} \pi^\beta [\![\kappa_b + \epsilon]\!] \leq 1, \qquad \Delta_{[\![\kappa_b+\epsilon]\!]} \left(\frac{F^a(k)}{p^b_k} \pi^\beta_k \right) \leq 0.$$

and similarly for asks. As $N \to \infty$, the difference equation converges to the ODE (5a) *with initial conditions given by* (5b)*.*

Note that there are intrinsically two sets of thresholds here: the κ_b and κ_a that correspond to a LOB with a finite starting state, and the thresholds of the LOB with an infinite starting state. We treat κ_b, κ_a as coming from the LOB with a finite starting state, so that they do not depend on N. For large N, the two sets of thresholds will be close to each other (this follows from the Lipschitz property of Lemma 3.1, and the characterization of κ_b in (3b)).

Proof. The proof proceeds in stages.

Stage 0. Let x_0 be given by $F^a(x_0) = F^b(\kappa_a) - F^b(x_0)$, and let y_0 be given by $1 - F^b(x_0) = (1 - F^a(\kappa_b)) - (1 - F^a(y_0))$. Equivalently, $F^a(x_0) + F^b(x_0) = 2x_0 = F^b(\kappa_a)$, so $x_0 = \frac{1}{2} F^b(\kappa_a)$, and $1 - y_0 = \frac{1}{2}(1 - F^a(\kappa_b))$.

Claim 0.1: $\kappa_b \leq x_0 < y_0 \leq \kappa_a$.

Proof: Note that $F^a(\kappa_b)$ is a lower bound on the rate of bid departure from the Markov chain when there are any bids present, while $F^b(\kappa_b) - F^b(\kappa_a)$ is an upper bound on the rate of bid arrival. Consequently, if $F^a(\kappa_b) > F^b(\kappa_b) - F^b(\kappa_a)$, then the number of bids on the entire interval (κ_b, κ_a) would be stochastically bounded, whereas it should scale as a random walk. A similar argument gives $y_0 \leq \kappa_a$. Finally, to see $x_0 < y_0$, recall from Proposition 4.2 that $\kappa_b < \kappa_a$, which implies $\frac{1}{2} F^b(\kappa_a) < \frac{1}{2}(1 + F^a(\kappa_b))$. □

Claim 0.2: There exists $T_0 = T_0(M)$ such that for all times $t \geq T_0$ and all fluid models, $\sum_{k=\llbracket x_0 \rrbracket+1}^{\llbracket y_0 \rrbracket-1}(q_k^b(t)+q_k^a(t)) = 0$.

Proof: Since these processes are absolutely continuous and nonnegative, it suffices to show that whenever there are any fluid orders in the interval (and all the derivatives are defined), the fluid number of orders in the interval decreases at a rate bounded below. By (7f) and (7g), we see that for $Q(t) = \sum_{\llbracket x_0 \rrbracket+1 \leq k \leq \llbracket \kappa_a \rrbracket-1} q_k^b(t)$,

$$Q'(t) \leq \begin{cases} 0, & Q(t) = 0 \\ \sum_{k=\llbracket x_0 \rrbracket+1}^{\llbracket \kappa_b \rrbracket-1} p_k^b - \sum_{k' \leq \llbracket x_0 \rrbracket+1} p_k^a < F^b(\kappa_b) - F^b(x_0) - F^a(x_0) - \epsilon, & Q(t) > 0. \end{cases}$$

Consequently, after a finite amount of time T_0^b, there will be no fluid bids in bins $\geq \llbracket x_0 \rrbracket + 1$. Similarly, after a finite amount of time T_0^a, there will be no fluid asks in bins $\leq \llbracket y_0 \rrbracket - 1$; we may take $T_0 = \max(T_0^b, T_0^a)$. $\square$

Claim 0.3: There exists $\epsilon_0 > 0$ such that for all times $t \geq T_0$ and all fluid models, $\sum_{k \leq \llbracket x_0 \rrbracket} \pi_k^\beta(t) \geq \epsilon_0$ and $\sum_{k \geq \llbracket y_0 \rrbracket} \pi_k^\alpha(t) \geq \epsilon_0$. (This result requires bins to be sufficiently small.)

Proof: Note that equations (7f) and (7g) hold at all times, even when there are no fluid orders in the bin; thus, for $t \geq T_0$ and all $k \in [\llbracket x_0 \rrbracket + 1, \llbracket y_0 \rrbracket - 1]$ we have

$$p_k^b \sum_{k'>k} \pi_{k'}^\alpha(t) = \pi_k^\beta(t) \sum_{k' \leq k} p_{k'}^a, \qquad p_k^a \sum_{k'<k} \pi_{k'}^\beta(t) = \pi_k^\alpha(t) \sum_{k' \geq k} p_{k'}^b.$$

Omitting the dependence on t for clarity, these equations, together with the observation that $\sum_k \pi_k^\alpha = \sum_k \pi_k^\beta = 1$, can be rearranged to give two decoupled second-order difference equations for π_k^α and π_k^β, as follows. Here, the operator Δ_k is given by $\Delta_k(f) = f_{k+1} - f_k$ for any sequence indexed by k, and we abuse notation to write $F^a(k) = \sum_{k' \leq k} p_{k'}^a$ and similarly for $F^b(k)$.

$$\Delta_k \left(\frac{1 - F^b(k)}{p_{k+1}^a} \cdot \Delta_k \left(\frac{F^a(k)}{p_k^b} \pi_k^\beta \right) \right) = \pi_{k+1}^\beta, \qquad \llbracket x_0 \rrbracket + 1 \leq k \leq \llbracket y_0 \rrbracket - 1. \tag{8a}$$

(There is a corresponding equation for π^a, of course.)

If we had two initial conditions for this second-order difference equation, we would be able to solve it. Unfortunately, in general we do not have such initial conditions, but we have bounds on them, namely

$$\frac{F^a(\llbracket x_0 \rrbracket)}{p_{\llbracket x_0 \rrbracket}^b} \pi^\beta \llbracket x_0 \rrbracket \leq 1, \qquad \Delta_{\llbracket x_0 \rrbracket} \left(\frac{F^a(k)}{p_k^b} \pi_k^\beta \right) \leq 0. \tag{8b}$$

These inequalities would hold with equality in a different limit order book $\tilde{\mathcal{L}}_0$, in which we assign the same low price to all the bins up through $\llbracket x_0 \rrbracket + 1$, and the same high price to all the bins from $\llbracket y_0 \rrbracket - 1$ up. (We nonetheless keep track the bins containing the highest bid and lowest ask of $\tilde{\mathcal{L}}$.) Corollary 4.4 shows that the solutions to (8) on $\llbracket x_0 \rrbracket + 1 \leq k \leq \llbracket y_0 \rrbracket - 1$ are bounded from above by the solution for $\tilde{\mathcal{L}}$. (The result is in continuous space, but the arguments work just as well for difference equations.) We refer to the solution for $\tilde{\mathcal{L}}$ as $\tilde{\pi}^\beta$ (and $\tilde{\pi}^\alpha$).

We now have

$$\sum_{k \leq \llbracket y_0 \rrbracket - 1} \pi_k^\beta \leq \sum_{k=\llbracket x_0 \rrbracket+1}^{\llbracket y_0 \rrbracket-1} \tilde{\pi}_k^\beta + \sum_{k=\llbracket y_0 \rrbracket}^{\llbracket \kappa_a \rrbracket-1} \frac{p_k^b}{\sum_{k' \leq k} p_{k'}^a}. \tag{9}$$

Notice that $\tilde{\pi}^\beta$ must satisfy $\tilde{\pi}_k^\beta = (F^a(k))^{-1} p_k^b$ for $[\tilde{\kappa}_b] + 1 \leq k \leq \llbracket x_0 \rrbracket$, as bids will not be queueing in those bins. Consequently, for the first term in the right-hand side of (9) we have

$$\sum_{k=\llbracket x_0 \rrbracket+1}^{\llbracket y_0 \rrbracket-1} \tilde{\pi}_k^\beta \leq 1 - \sum_{k=[\tilde{\kappa}_b]+1}^{\llbracket x_0 \rrbracket} \frac{p_k^b}{F^a(k)} \leq 1 - \sum_{k=\llbracket y_0 \rrbracket}^{\llbracket \kappa_a \rrbracket-1} \frac{p_k^b}{F^a(k)} - \epsilon_0,$$

where the last inequality holds provided the bins are sufficiently narrow. Indeed, notice that $x_0 - \tilde{\kappa}_b > x_0 - \kappa_b = \kappa_a - y_0$ (from monotonicity of $\mathcal{L}$ vs. $\tilde{\mathcal{L}}$ and symmetry), the denominator is increasing in k, and the bid arrival density decreases with translation to the right. It is possible for finitely many bins that the sums are empty, but if the bins are narrow enough this will not occur.

Stage 1. We now let x_1, y_1 be defined by $F^b(x_0) - F^b(x_1) = \epsilon_0 F^a(x_1)$ and $F^a(y_1) - F^a(y_0) = \epsilon_0(1 - F^b(y_1))$. Similarly to the argument for Stage 0, there exists a time T_1 such that for all $t \geq T_1$ there will be

no fluid queues on $[\![x_1]\!]+1, [\![y_1]\!]-1]$. Indeed, if there are fluid bids in the interval $[[\![x_1]\!]+1, [\![x_0]\!]]$, then whenever the highest bid is below $[\![x_0]\!]$ it is in fact in this interval; the defining inequality then means that the fluid amount of bids in this interval decreases, and similarly for asks.

Next, we use the difference equation description on $[[\![x_1]\!]+1, [\![y_0]\!]-1]$ to show that after T_1, the highest bid spends at least $\epsilon_1 > 0$ of its time below x_1. This will require comparison against a different restricted LOB $\tilde{\mathcal{L}}_1$, where we merge all prices up to $[\![x_1]\!]+1$ and from $[\![y_1]\!]-1$.

Subsequent stages. We can now construct a nested sequence of intervals $\ldots < x_2 < x_1 < x_0 < y_0 < y_1 < y_2 < \ldots$, where the inequalities are strict provided bins are narrow enough. It remains to show that $\lim_{k\to\infty, N\to\infty} x_k = \kappa_b$ and $\lim_{k\to\infty, N\to\infty} y_k = \kappa_a$. (Note that $N \to \infty$, i.e. thinner bins, is certainly necessary for this to hold!)

This result follows from the fact that ϵ_i are bounded below:

$$\epsilon_i \geq \sum_{k=[\tilde{\kappa}_b]+1}^{[\![x_i]\!]} \frac{p_k^b}{\sum_{k'\leq k} p_{k'}^a} - \sum_{k=[\![y_i]\!]}^{[\![\kappa_a]\!]-1} \frac{p_k^b}{\sum_{k'\leq k} p_{k'}^a} \geq \left(\frac{1}{F^a([\![x_i]\!])} - \frac{1}{F^a([\![y_i]\!])}\right)\left(F^b([\![x_i]\!]) - F^b([\tilde{\kappa}_b]+1)\right).$$

It is clear that as long as x_i is bounded away from κ_b (and bin widths are small enough), this will be bounded below, and therefore $x_i - x_{i+1}$ and $y_{i+1} - y_i$ will be bounded below.

Convergence to ODE. The convergence of difference equations to ODE is standard. The argument above gives an inequality for the initial conditions, but note that as we approach κ_b the initial conditions become exact. Indeed,

$$F^a(\kappa_b+\epsilon)\varpi^b(\kappa_b+\epsilon) = \int_{\kappa_b+\epsilon}^1 \varpi^a(x) f^a(x) dx \to 1,$$

since the lowest ask will never be below κ_b. Also,

$$(F^a(x)\varpi^b(x))'|_{x=\kappa_b+\epsilon} = -\varpi^a(\kappa_b+\epsilon) = -(1-F^b(\kappa_b+\epsilon))^{-1}\int_0^{\kappa_b+\epsilon} \varpi^b(x) f^b(x) dx \to 0,$$

since the highest bid density is bounded. □

Putting this result together with Proposition 4.2 shows that, for symmetric distributions p^b, p^a with p^b decreasing, the fluid limits $\pi_k^\beta(t)/p_k^b$, $\pi_k^\alpha(t)/p_k^a$ will approach, as $t \to \infty$ and $N \to \infty$, the solution of the ODE (5), uniformly on compact subsets of (κ_b, κ_a).

It remains to show that stability of fluid limits implies positive recurrence of the Markov chain.

Lemma 4.7 (Fluid stability and positive recurrence)**.** *Consider a LOB satisfying the assumptions of Theorem 4.6. Suppose that on some interval of bins $k_0 \leq k \leq k_1$, all fluid limits with initial state bounded above by 1 satisfy the following: there exists a time T depending on $\{M, p_k^a, p_k^b, \text{bin widths}\}$, such that for all times $t \geq T$,*

$$q_k^b(t) = 0, \qquad q_k^a(t) = 0, \qquad k_0 \leq k \leq k_1,\ t \geq T.$$

Consider a limit order book $\tilde{\mathcal{L}}$ started with infinitely many bids in bin $k_0 - 1$ and infinitely many asks in bin k_1; its state is described by the Markov chain of queue sizes in bins $k_0 \leq k \leq k_1$. This Markov chain associated to $\tilde{\mathcal{L}}$ is positive recurrent.

Proof. To go between fluid stability and positive recurrence, we use the multiplicative Foster's criterion [13, Theorem 13.0.1]. Let

$$Q(t) = \left\|(Q_k^b(t), Q_k^a(t))_{k_0\leq k\leq k_1}\right\|,$$

and let $C > M$ be sufficiently large. Let $Q(0) = q > C$, and consider the fluid scaling $\overline{Q}_k^{a,b}(t) = q^{-1}Q_k^{a,b}(qt)$. By Theorem 4.5, if C and hence q is large enough, there exists a fluid limit $(q_k^a(t), q_k^b(t), \tau_k^\alpha(t), \tau_k^\beta(t))_{k_0\leq k\leq k_1}$ satisfying $\left\|q_k^a(t), q_k^b(t)\right\| = 1$, such that $\mathbb{P}(\left\|\overline{Q}_k^a(t) - q_k^a(t), \overline{Q}_k^b(t) - q_k^b(t)\right\| > \epsilon) \leq \epsilon$ for all $t \in [0,T]$. In particular, $\mathbb{P}(\|Q_k^a(qT), Q_k^a(qT)\| > \epsilon q) < \epsilon$. Note further that $\left\|Q_k^a(qT), Q_k^b(qT)\right\| \leq qT$ simply because orders arrive deterministically at rate 1. Thus, we conclude

$$\mathbb{E}_q[\left\|Q_k^a(qT), Q_k^b(qT)\right\|] \leq \epsilon(1+T)q.$$

Choosing $\epsilon < (1+T)^{-1}$ completes the proof. □

4.4. **General order price distributions.** It remains to remove the extra conditions (symmetric and decreasing) on the order price distributions, and finish the argument for continuous limit order books. This requires two observations:

(1) Recall that a continuous LOB could be bounded by two discrete non-strict LOBs with different arrival price distributions (in one of them, we shift all arriving bids one bin to the left). This shifted arrival distribution no longer satisfies the absolute continuity conditions, but nevertheless, Lemma 3.1 shows that all of the above fluid-scaled arguments work for it as bin size shrinks to 0. Specifically, we model the bid arrivals as shifting the *rightmost* bin of bids all the way to the left, and then the difference between the two books is at most two bins' worth of arrivals over the fluid time interval $[0, T]$, which will be small provided bins are narrow. This allows us to conclude the positive recurrence of a continuous LOB with infinitely many bids at price $\mathcal{P}(\kappa_b)+\epsilon$ and infinitely many asks at price $\mathcal{P}(\kappa_a)-\epsilon$, provided f^a, f^b satisfy absolute continuity assumptions, are symmetric, and f^b is decreasing.

(2) Recall that replacing the bid arrival price distribution by another distribution with stochastically higher prices, and/or replacing he ask arrival price distribution by another distribution with stochastically lower prices, results in fewer orders in a book. In particular, if we have shown the positive recurrence of an LOB with an infinite supply of bids at price p and asks at price q with a particular arrival distribution, the LOB will remain positive recurrent when we switch to an arrival price distribution with bids further right, and asks further left. Notice that as long as there are bids in the interval (p, q), they evolve on that interval identically whether or not there is an infinite supply of bids at p; and similarly for asks. This can be used to show that fluid limits drain in the new LOB on the interval (p, q).

In the new LOB with the shifted price distribution, (p, q) may not be close to $(\tilde\kappa_b, \tilde\kappa_a)$, so we will be wanting to extend the interval, as in Claim 0.3 of Theorem 4.6. The argument there does not use the full extent of the symmetry and monotonicity conditions; they are only used to prove the inequality

$$\sum_{k=[\![\hat\kappa_b]\!]+1}^{[\![p]\!]} \frac{p_k^b}{F^a(k)} \geq \sum_{k=[\![q]\!]}^{[\![\tilde\kappa_a]\!]-1} \frac{p_k^b}{F^a(k)} + \epsilon$$

for some $\epsilon > 0$. (Here, we had $\hat\kappa_b < \tilde\kappa_b$.) For this inequality to hold, it is entirely sufficient to only assume $p - \tilde\kappa_b = \tilde\kappa_a - q$ and $f^b(x) > f^b(y)$ if $x \in (\tilde\kappa_b, p)$ and $y \in (q, \tilde\kappa_a)$; or more generally, that

$$\int_{\hat\kappa_b}^{p} \frac{f^b(x)}{F^a(x)}dx \geq \int_{q}^{\tilde\kappa_a} \frac{f^b(x)}{F^a(x)}dx + \epsilon.$$

(So we may have $\tilde\kappa_a - q \neq p - \tilde\kappa_b$ as long as this inequality still holds.)

Consequently, for general (f^b, f^a) satisfying the absolute continuity conditions, we begin by finding f_0^b, f_0^a with $F_0^b \leq F^b$, $F_0^a \geq F^a$ which are symmetric and for which f_0^b is decreasing. We use Theorem 4.6 to show that fluid limits drain for f_0^b, f_0^a (and hence for (f^b, f^a)) on an interval $(\kappa_{b,0}, \kappa_{a,0})$. We then modify the distributions on $(\kappa_{b,0}, \kappa_{a,0})$ to find (f_1^b, f_1^a) satisfying the absolute continuity conditions, for which $F_0^b \leq F_1^b \leq F^b$, $F^a \leq F_1^a \leq F_0^a$. We already know from monotonicity that fluid limits will drain for these distributions on $(\kappa_{b,0}, \kappa_{a,0})$, and we use the inequality for $p \leq \kappa_{b,0}$ and $q \geq \kappa_{a,0}$ to extend fluid stability to the bigger interval $(\kappa_{b,1}, \kappa_{a,1})$. We repeat the process until the interval $(\kappa_{b,n}, \kappa_{a,n})$ approaches the entire interval (κ_b, κ_a) for the original pair of distributions (f^b, f^a).

Notice that all that really matters for the thresholds κ_b and κ_a of a LOB is $F^{a,b}(x), \kappa_b \leq x \leq \kappa_a$; it is immaterial what f^b and f^a do outside of those intervals, so long as they integrate to the correct amounts. Consequently, if $\kappa_{b,n} > \kappa_b + \epsilon$, it must be that $F_n^b < F^b$ or $F_n^a > F^a$ somewhere on $[\kappa_{b,n}, \kappa_{a,n}]$, which means that the process won't get "stuck" until $\kappa_{b,n} \searrow \kappa_b$ and $\kappa_{a,n} \nearrow \kappa_a$.

5. Discussion

In this section we discuss several applications of our earlier methods and results. We begin with a discussion of market orders and then consider various simple trading strategies.

5.1. **Market orders.** The orders we have considered so far, each with a price attached, are called limit orders. Suppose that, in addition to limit orders, there are also *market orders* which request to be fulfilled immediately at the best available price. Suppose that limit order bids and asks arrive as independent Poisson processes of rates ν_b, ν_a respectively; and that the prices associated with limit order bids, respectively asks, are independent identically distributed random variables with density $f_b(x)$, respectively $f_a(x)$. Without loss of generality we may assume that $x \in (0,1)$. In addition suppose that there are independent Poisson arrival streams of market order bids and asks of rates μ_b, μ_a respectively. Then these correspond to extreme limit orders: we simply associate a price 1 or 0 with a market bid or market ask respectively.

Note that, in addition to market orders, we have also allowed an asymmetry in arrival rates between bid and ask orders. The intuition behind equations (1) leads to the generalization

$$\nu_b f_b(x) \int_x^{\kappa_a} \pi_a(y)dy = \pi_b(x) \left(\mu_a + \nu_a \int_0^x f_a(y)dy \right) \tag{10a}$$

$$\nu_a f_a(x) \int_{\kappa_b}^x \pi_b(y)dy = \pi_a(x) \left(\nu_b \int_x^1 f_b(y)dy + \mu_b \right) \tag{10b}$$

although now the existence of a solution to these equations satisfying the required boundary conditions is not assured, and the deduction of the recurrence properties necessary for an interpretation of $\pi_b(x), \pi_a(x)$ as limiting densities may fail. To illustrate some of the possibilities we shall look in detail at a simple example.

Suppose $f_a(x) = f_b(x) = 1, x \in (0,1)$, $\nu_a = \nu_b = 1 - \lambda$ and $\mu_a = \mu_b = \lambda$. Thus a proportion λ of all orders are market orders. Use the notation $\pi_b(\lambda; x)$ for the solution to equations (10) satisfying the required boundary conditions in this example. Then this solution is

$$\pi_b(\lambda; x) = \pi_b \left(\frac{1+\lambda}{1-\lambda} x - \frac{\lambda}{1-\lambda} \right) \tag{11}$$

where $\pi_b(.)$ is the earlier solution (2) provided $\lambda < w \approx 0.278$, the unique solution of $we^w = e^{-1}$. Indeed, provided $\lambda < w$ the model is simply a rescaled version of the earlier model with distribution (11) having a support increased from $(\kappa, 1-\kappa)$ to the wider interval $(\kappa(\lambda), 1-\kappa(\lambda))$ where

$$\kappa(\lambda) = \frac{1+\lambda}{1-\lambda} \cdot \frac{w}{1+w} - \frac{\lambda}{1-\lambda}.$$

The inclusion of market orders in the model causes the price distributions to have atoms and not to be absolutely continuous with respect to each other; but nevertheless the analysis of earlier sections continues to apply since the market orders arrive outside of the range $(\kappa(\lambda), 1-\kappa(\lambda))$.

Next we explore this example as $\lambda \uparrow w$ and the support becomes the entire interval $(0,1)$. In our model a market order bid, respectively ask, which arrives when there are *no* ask, respectively bid, limit orders in the order book waits until it can be matched. When $\lambda < w$ there is a finite (random) time after which the order book always contains limit orders of both types and no market orders of either type and hence the analysis of previous sections applies. But if $\lambda > w$ then infinitely often there will be no asks in the order book and infinitely often there will be no bids in the order book, with probability 1. Now the difference between the number of bid and ask orders in the limit book is a simple symmetric random walk and hence null recurrent. There will infinitely often be periods when the order book contains limit orders of both types and no market orders of either type, but such states cannot be positive recurrent.

In the model described above an arriving market order which cannot be matched immediately must wait until it can be matched. If instead such orders are lost then we obtain a model which can be analyzed by the methods in Section 5.2.1: namely, we start the LOB with an infinite bid order at 0 and an infinite ask order at 1.

5.2. **Trading strategies.** Next we consider a few simple strategies that can be analyzed using our model. For simplicity, we present the results for the case when the bid and ask price distributions are equal and uniform on $(0,1)$, but the analysis easily extends to other arrival distributions. Recall that the limiting densities of the rightmost bid and leftmost ask were determined in Corollary 2.3.

5.2.1. *Market making.* We begin by considering a single market maker who places an infinite number of bid, respectively ask, orders at p respectively $q = 1 - p$, where $\kappa_b < p < q < \kappa_a$. Thus whenever q is the lowest ask price, the trader obtains all bids that arrive at prices above q, and whenever p is the highest bid price, she obtains all asks that arrive at prices below p, making a profit of $q - p$ per bid–ask pair so acquired. Call the orders placed by the trader *artificial*, to distinguish them from the *natural* orders. Notice that the rate at which the trader is able to match her orders is proportional to p times the probability that the rightmost bid is exactly p.

Note that placing an infinite supply of bids at a level below κ has no effect on the evolution of the LOB. For $p > \kappa \approx 0.217$ no ask is accepted at a price less than p, and there will be a positive probability that the rightmost bid is exactly p. This probability mass is exactly the probability that the rightmost bid is p or less in a model with no asks arriving to the left of p and with no artificial bids. For this model the rightmost natural bid has density $1/x$ on $[\kappa_b, p)$, and $C\left(\frac{1}{x} + \log\frac{1-x}{x}\right)$ on $[p, q]$, for C such that the resulting density is continuous. This allows us to find κ_b as

$$\kappa_b = \frac{p}{e}\left(\frac{1-p}{p}\right)^C$$

and to deduce that the rightmost natural bid has density

$$\varpi^b(x) = \begin{cases} \dfrac{1}{x}, & \dfrac{p}{e}\left(\dfrac{1-p}{p}\right)^C \le x \le p; \\ C\left(\dfrac{1}{x} + \log\dfrac{1-x}{x}\right), & p \le x \le q; \end{cases}$$

where $C = (1 + p\log((1-p)/p))^{-1}$. The probability the rightmost natural bid is p or less is thus $1 - C\log((1-p)/p)$, and this is therefore the probability that the rightmost bid is exactly p in the model with infinitely many artificial bids at p and infinitely many artificial asks at $q = 1 - p$, where $\kappa < p < 1/2 < q$.

Thus to maximize the profit rate we need to solve the optimization problem

$$\begin{aligned} &\text{maximize} \quad (1-2p)p(1-C) \quad \text{where} \quad C = (1-p)\log\frac{1-p}{p} \\ &\text{subject to} \quad p \in [\kappa, 1/2]. \end{aligned}$$

The maximum is attained at $p \approx 0.369$, and gives a profit rate of ≈ 0.064.

5.2.2. *Sniping.* We next consider a trader with a sniping strategy: the trader immediately buys every bid that joins the LOB queue at price above q, and every ask that joins the LOB queue at price below p (with $q = 1 - p$ still). There is a twofold difference between this model and the market making model: here, the trader has lower priority than the orders already in the queue, but she obtains a better price for the orders that she does manage to buy.

The effect on the LOB of the sniping strategy is to ensure no queued bids above q and no queued asks below p; for $p < q$, the set of bids and asks on (p, q) is the same in the sniping and the market making model. But it also makes sense to consider the sniping strategy with $p > q$, when it ensures that there are no queued orders of any kind in the interval (q, p): they are all sniped up by the trader. (An ask arriving at price $a \in (q, p)$ cannot be matched with a queued bid, because there are no queued bids above q.) In particular, the trader makes a net profit of zero on the orders in (q, p); the point of sniping them is to increase the probability of being able to buy a bid at a high price.

Summarizing, if $p > q$ then the LOB has no queued orders between p and q and rightmost bid has density $1/x$ on $[\kappa_b, q)$, where $\kappa_b = q/e$. Notice that the distribution of the rightmost bid stochastically decreases as q decreases, hence the probability of acquiring an ask at low price $a < 1/2$ increases as q decreases. This shows that profit from sniping bids above q and asks below p for $p > 1/2$ is strictly higher than the profit from sniping bids above p and asks below q. Thus, it suffices to consider the case of $p > 1/2 > q$. We thus

solve

$$\begin{aligned} &\text{maximize} && \int_p^{\kappa_a} (2b-1)\log\frac{1-b}{1-\kappa_a}\,db \\ &\text{where} && \kappa_a = 1 - \frac{1-p}{e} \\ &\text{subject to} && p \in [1/2, 1]. \end{aligned}$$

The maximum is attained at $p = (e^2 - e + 1)/(e^2 + 1) \approx 0.676$ and gives a profit rate of ≈ 0.060. Perhaps surprisingly, this is lower than the optimized profit rate with a market making strategy.

Figure 2 presents a comparison between the profit rates from the market making and sniping strategies, as a function of p (which, recall, is the price below which the trader would like all asks) – for completeness, $p < 1/2$ is included for the sniping strategy as well. Notice that the optimum for the market making strategy has $p < 0.5$, while the optimum for the sniping strategy has $p > 0.5$.

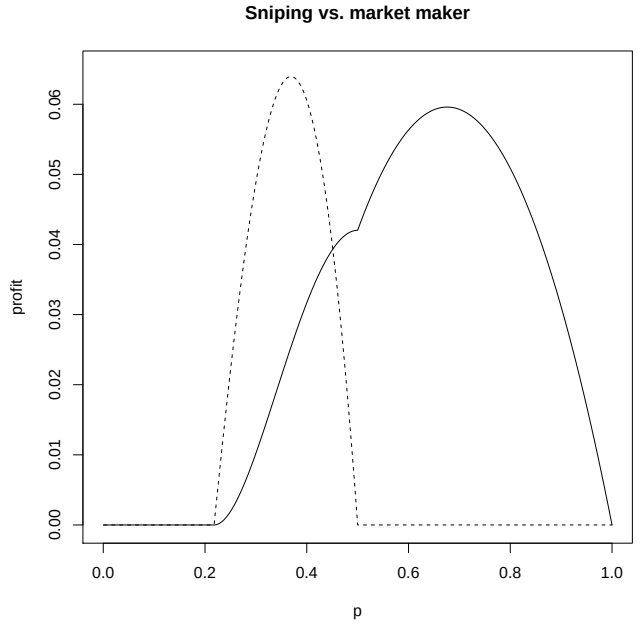

FIGURE 2. Profit from sniping and market making strategies. Solid line is the sniping strategy, dashed line is the market maker strategy. (Sniping with $p < 1/2$ is shown for completeness; as argued in the text, it does not maximize the profit.)

5.2.3. *A mixed strategy.* It is possible to consider a mixture of the above strategies: the trader places an infinite supply of bids at P (thus acquiring all asks that arrive below P whenever P is the highest bid price), but in addition attempts to snipe up all the additional asks that land at prices $x < p$. We assume the trader gets the best of the two possible prices when both p and P are larger than the price of the arriving ask. There are several possible cases corresponding to the relative arrangement of p, P, and $1/2$:

(1) If $p < P$ (this means that there are no additional asks to snipe up), this degenerates to the market maker strategy, with a profit of $(1 - 2P)$ per bid–ask pair bought, with pairs bought at rate $P\log(P/\kappa_b)$. (The probability of the highest natural bid being below P is $\log(P/\kappa_b)$; when it is there, asks arrive at prices below P at rate P.) Clearly, one wants $P < 1/2$ in this case, otherwise the profit is negative, so we can write this case as $p < P < 1/2$.
(2) If $P < p < 1/2$, then one gets additional asks at price x at rate $\log(x/\kappa_b)$, for a profit of $(1 - 2x)$, for all x from P to p.
(3) If $P < 1/2 < p$, there are two further cases: we may have $P < 1 - p$ or $P > 1 - p$.

(a) If $P < 1-p < 1/2 < p < 1-P$, then the trader snipes all orders between $1-p$ and p for a net profit of 0. Profit $(1-2P)$ from a bid–ask pair matching the infinite orders is generated at rate $P\log((1-p)/\kappa_b)$, and profit $1-2x$, $P \le x \le 1-p$, from sniping is generated at rate $1+\log(x/\kappa_b)$. Note that here the bid density is $1/x$ on $(\kappa_b, 1-p]$, so $\kappa_b = (1-p)/e$.

(b) If $1-p < P < 1/2 < 1-P < p$, then P is always the best bid, which means that the trader gets all the asks that arrive below P, generating profit at rate $(1-2P)P$. Orders arriving between P and $1-P$ cancel each other, and all the asks arriving between $1-P$ and p are bought up for a loss (negative profit) of $(1-2x)$.

(4) Finally, the case $P > 1/2$ is silly, because every bid–ask pair bought will be bought at a loss.

Figure 3 shows the profit for the two-parameter space. The largest profit is obtained when $P = 1-p = 1/4$, and the profit is then acquired at rate $1/8 = 0.125$. This corresponds to the trader placing an infinite bid order at $1/4$ (thus buying all asks that arrive with price below $1/4$ for $1/4$), an infinite ask order at $3/4$, and sniping up all orders that join the LOB at prices between $1/4$ and $3/4$.

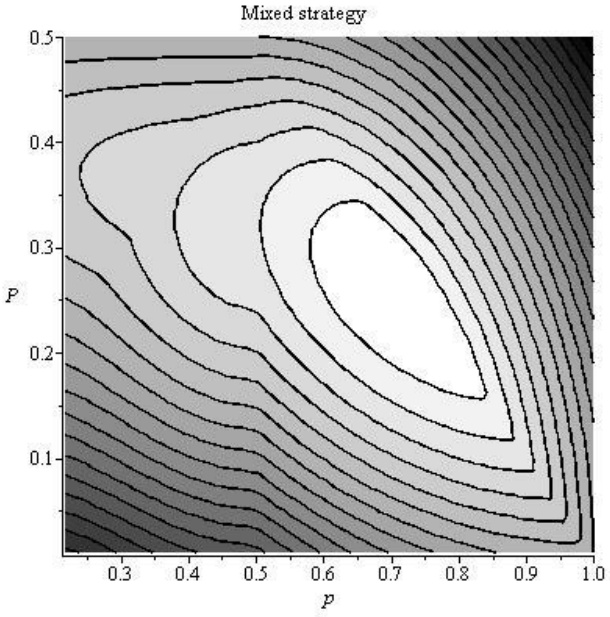

FIGURE 3. Profit rate from the mixed strategy, as a function of sniping threshold p and infinite bid order location P.

5.3. **Nash equilibrium.** Finally we analyze the situation that arises when multiple traders compete. There is clearly an advantage for a trader who can snipe an order more quickly than the other traders. It has been argued that competition on speed is wasteful [3], and there are proposals to encourage traders to compete on price, rather than speed, as for example in the proposal [4] where a market continuous in time is replaced with frequent batch auctions, held perhaps several times a second. We shall explore the consequences of competition on price between high-frequency traders who can all react at the same speed to a new order entering the LOB.

We will look for a Nash equilibrium between traders each using the mixed strategies of Section 5.2.3, since these are easy to implement and analyze. Notice first of all that no Nash equilibrium strategy can involve any compulsory trades that incur a loss, since it is always optimal for a single trader to opt out of those. In addition, the Nash equilibrium strategy cannot involve infinite orders: if those make a profit, then it is optimal for a trader to improve the price of her own infinite bid order infinitesimally, thus gaining all the profit from those orders for herself alone. There is the possibility of two infinite orders at $1/2$, but it incurs no profit for anyone in the market.

Thus, we are reduced to considering the space of sniping strategies, in which one only snipes at asks with prices below $1/2$ (or bids at prices above $1/2$). In this case, it is clearly preferable for each trader to snipe at all such orders (getting the order with the same probability as each of the other traders), and we conclude that the Nash equilibrium is for all traders to snipe at all asks below $1/2$ and at all bids with prices above $1/2$. The rightmost bid will then have density $1/x$ on $(1/(2e), 1/2)$. This results in a combined profit rate $1/(2e) - (1+e^2)/(8e^2) \approx 0.042$. Thus price competition between traders has decreased their combined profit rate from 0.125 to 0.042.

Next we comment on the impact of traders on the bid-ask spread. The mean of the distribution (2) can be calculated and is simply $(1-\kappa)/2$. Thus without traders the mean spread between the highest bid and the lowest ask in the LOB is $\kappa \approx 0.218$, while the maximum spread is $1 - 2\kappa \approx 0.564$. At the Nash equilibrium between traders both are increased, the mean spread to $1/e \approx 0.368$ and the maximum spread to $1 - 1/e \approx 0.632$. These calculations are for a uniform price distribution, but provided the arriving bid and ask prices distributions are identical the results give the mean and maximum spreads in terms of percentiles of that distribution.

As a final remark we comment on the inventory of traders under the Nash equilibrium described above. Observe that the LOB below $1/2$ evolves independently of the LOB above $1/2$, and both processes are positive recurrent. Consider the net position of the traders collectively, that is all the bids they have matched minus all the asks they have matched, observed at those times when the LOB is empty. This evolves as a symmetric random walk, and is null recurrent. Slight variations of the traders strategies would moderate this conclusion: for example, a trader might refrain from sniping bids close enough to $1/2$ when her net position is large. And of course such variations will be essential over longer time-scales than those considered in this paper where the arrival price distributions may vary.

References

[1] I. Adan and G. Weiss. Exact FCFS matching rates for two infinite multitype sequences. *Operations Research*, 60:475–489, 2012.

[2] Maury Bramson. *Stability and Heavy Traffic Limits for Queueing Networks: St. Flour Lectures Notes*. Springer, 2006. `http://www.math.duke.edu/~rtd/CPSS2007/Bramson.pdf`.

[3] E. Budish, P. Cramton, and J. Shim. The high-frequency trading arms race: Frequent batch auctions as a market design response. `http://faculty.chicagobooth.edu/eric.budish/research/HFT-FrequentBatchAuctions.pdf`, 2013.

[4] E. Budish, P. Cramton, and J. Shim. Implementation details for frequent batch auctions: Slowing down markets to the blink of an eye. *American Economic Review*, 104:418–424, 2014.

[5] R. Cont and A. de Larrard. Price dynamics in a markovian limit order book market. *SIAM Journal of Financial Mathematics*, 4:1–25, 2013.

[6] R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. *Operations Research*, 58:549–563, 2010.

[7] D. Easley, M. Lopez de Prado, and M. O'Hara. The volume clock: Insights into the high frequency paradigm. *The Journal of Portfolio Management*, 39:19–29, 2012.

[8] David Gamarnik and D. Katz. Stability of Skorokhod problem is undecidable. Submitted, 2010. `arXiv:1007.1694v1`.

[9] X. Gao, J. G. Dai, A. B. Dieker, and S. J. Deng. Hydrodynamic limit of order book dynamics. `http://arxiv.org/pdf/1411.7502.pdf`, 2014.

[10] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. *Quantitative Finance*, 13:1709–1742, 2013.

[11] David Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society*, 13(2):151–185, 1951.

[12] A. Lachapelle, J.-M. Lasry, C.-A. Lehalle, and P.-L. Lions. Efficiency of the price formation process in presence of high frequency participants: a mean field game analysis. `http://arxiv.org/pdf/1305.6323v3.pdf`, 2014.

[13] Sean Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, second edition, 2009.

[14] Ioanid Roşu. A dynamic model of the limit order book. *Review of Financial Studies*, 22:4601–4641, 2009.

[15] Alexander L. Stolyar and Elena Yudovina. Systems with large flexible server pools: Instability of "natural" load balancing. *Annals of Applied Probability*, 23:2099–2138, 2013.

[16] S. A. Zenios, G. M. Chertow, and L. M. Wein. Dynamic allocation of kidneys to candidates on the transplant waiting list. *Operations Research*, 48:549–569, 2000.

Frank Kelly, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Rd, Cambridge CB3 0WB, United Kingdom, E-mail: `fpk1@cam.ac.uk`, Elena Yudovina, Department of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church St. S.E., Minneapolis, MN 55455, E-mail: `eyudovin@umn.edu`