5

Single-Queue Dynamics: Simple Models

Nothing is more practical than a good theory.

(L. Boltzmann)

Modelling the full dynamics of an LOB is a complicated task. As we discussed in Chapter 3, limit orders can be submitted or cancelled at a wide range of different prices, and can also be matched to incoming market orders. Limit orders of many different sizes often reside at the same price level, where they queue according to a specified priority system (see Section 3.2.1). The arrival and cancellation rates of these orders also depend on the state of the LOB, which induces a feedback loop between order flow and liquidity and thereby further complicates the problem. Due to the large number of traders active in some markets, and given that each such trader can own many different limit orders at many different prices, even keeping track of an LOB's temporal evolution is certainly a challenge.

Despite these difficulties, there are many clear benefits to developing and studying LOB models. For example, analysing the interactions between different types of orders can help to provide insight into how best to act in given market situations, how to design optimal execution strategies, and even how to address questions about market stability. Therefore, LOB modelling attracts a great deal of attention from practitioners, academics and regulators.

Throughout the next four chapters, we introduce and develop a framework for LOB modelling. In the present chapter, we begin by considering the core building block of our approach: the temporal evolution of a single queue of limit orders, using highly simplified models. In Chapter 6, we extend our analysis to incorporate several important empirical facts into our theoretical description of single queues. In Chapter 7, we consider the joint dynamics of the best bid- and ask-queues together, from both a theoretical and an empirical point of view. Finally, in Chapter 8, we discuss how to extend these models to describe the dynamics of a full LOB. In all of these chapters, we aim to derive several exact results within

the framework of simplified stochastic models, and approximate results for more realistic models calibrated to market data.

5.1 The Case for Stochastic Models

An immediate difficulty with modelling LOBs concerns the reasons that order flows exist at all. At one extreme lies the approach in which orders are assumed to be submitted by *perfectly rational* traders who attempt to maximise their utility by making trades in markets driven by exogenous information. This approach has been the starting point for many microstructural models within the economics community. However, this extreme assumption has come under scrutiny, both because such models are (despite their mathematical complexity) unable to reproduce many salient empirical stylised facts and because perfect rationality is difficult to reconcile with direct observations of the behaviour of individuals, who are known to be prone to a variety of behavioural and cognitive biases. Therefore, motivating order submissions in a framework of perfect rationality is at best difficult and at worst misleading.

At the other extreme lies an alternative approach, in which aggregated order flows are simply assumed to be governed by stochastic processes. Models that adopt this approach ignore the strategies employed by individual market participants, and instead regard order flow as random. Due to its exclusion of explicit strategic considerations, this approach is often called **zero-intelligence** (ZI) modelling. In a zero-intelligence LOB model, order flow can be regarded as a consequence of traders blindly following a set of stochastic rules without strategic considerations.

The purpose of ZI models is not to claim that real market participants act without intelligent decision making. Instead, these models serve to illustrate that in some cases, the influence of this intelligence may be secondary to the influence of the simple rules governing trade.

In many situations, ZI frameworks are too simplistic to reproduce the complex dynamics observed in real markets. In some cases, however, ZI models serve to illustrate how some seemingly non-trivial LOB properties or behaviours can actually emerge from the interactions of simple, stochastic order flows.

Much like perfect rationality, the assumptions inherent in ZI models are extreme simplifications that are inconsistent with some empirical observations. However, the ZI approach has the appeal of leading to quantifiable models without the need for auxiliary assumptions. Therefore, we adopt this approach for the simple benchmark LOB models that we consider in the following chapters. ZI models can often be extended to incorporate "boundedly rational" considerations, which serve to improve their predictive power by incorporating simple constraints that

ensure the behaviour of traders acting within them is not obviously irrational, or to remove obvious arbitrage opportunities.

5.2 Modelling an Order Queue

Given that an LOB consists of many different order queues at many different prices, a sensible starting point for developing a stochastic model of LOB dynamics is to first concentrate on just one such queue. This enables us to ignore the interactions (including any possible correlations in order flow) between the activity at different price levels. This is the approach that we take throughout this chapter. Specifically, we consider the temporal evolution of a single queue of orders that is subject to limit order arrivals, market order arrivals and cancellations.

Let $V(t)$ denote the total volume of the orders in the queue at time t. Because we consider only a single order queue, the models studied in this chapter can be used for both the bid-queue and for the ask-queue. Assume that:

- all orders are of unit size $\upsilon_0 = 1$;
- limit orders arrive at the queue as a Poisson process with rate λ;
- market orders arrive at the queue as a Poisson process with rate μ;
- limit orders in the queue are cancelled as a Poisson process with some state-dependent rate (which we will specify in the following sections).

Despite the apparent simplicity of this modelling framework, we will see throughout the subsequent sections that even understanding the single-queue dynamics created by these three interacting order flows is far from trivial, and leads to useful insights about the volume dynamics at the best quotes.

5.3 The Simplest Model: Constant Cancellation Rate

To gain some intuition of the basic dynamics of the model in Section 5.2, we first consider the case in which the probability per unit time that exactly one order in a queue is cancelled is constant ν, independent of instantaneous queue length $V(t)$.

For $V \geq 1$, the model corresponds to the following stochastic evolution rules between times t and $t + \mathrm{d}t$:

$$\begin{cases} V \to V+1 & \text{with rate } \lambda \mathrm{d}t \quad \text{(deposition)}, \\ V \to V-1 & \text{with rate } (\mu+\nu)\mathrm{d}t \quad \text{(execution + cancellation)}. \end{cases} \tag{5.1}$$

Intuitively, our model regards the queue volume as a random walker with "position" V, with moves that occur as a Poisson process with rate $(\lambda + \mu + \nu)$. When a move occurs, the probability that it is an upward move is $\lambda/(\lambda+\mu+\nu)$ and the probability that it is a downward move is $(\mu+\nu)/(\lambda+\mu+\nu)$.

The case where the queue is empty requires further specification, because executions and cancellations are impossible when $V = 0$. To keep our calculations as simple as possible, we assume that as soon as the queue size depletes to zero, the queue is immediately replenished with some volume $V \geq 1$ of limit orders, chosen from a certain distribution $\varrho(V)$. We return to this discussion at many points throughout the chapter.

5.3.1 The Master Equation

To analyse our model in detail, we first write the master equation, which describes the temporal evolution of the probability $P(V,t)$ of observing a queue of length V at time t. Counting the different Poisson events with their respective probabilities, one readily obtains, for $V \geq 1$:

$$\frac{\partial P(V,t)}{\partial t} = -(\lambda+\mu+\nu)P(V,t) + \lambda P(V-1,t) + (\mu+\nu)P(V+1,t) + J(t)\varrho(V), \quad (5.2)$$

where the assumption that the queue replenishes as soon as it reaches $V = 0$ implies that $P(V = 0,t) = 0$. The probability per unit time of such depletion events is $J(t)$, which justifies the presence of the **reinjection current** (equal here to the **exit flux**) $J(t)\varrho(V)$, which represents replenished queues with initial volume V.

Given a dynamic equation such as in Equation (5.2), it is often desirable to seek a **stationary solution** $P_{\text{st.}}(V)$, for which

$$\frac{\partial P(V,t)}{\partial t} = 0.$$

In the stationary solution, the probability of observing a queue with a given length V does not evolve over time. In this case, it must also hold that $J(t)$ does not depend on time.

When $\lambda > \mu + \nu$, there cannot be any stationary state since the limit order arrival rate is larger than the departure rate (see Equation (5.6) below). In other words, there is a finite probability that the queue size grows unboundedly, without ever depleting to 0. This behaviour is clearly unrealistic for bid- and ask-queues, so we restrict our attention to the case where $\lambda \leq \mu + \nu$.

Let us first assume that the reinjection process is such that $\varrho(V) = 1$ for $V = V_0$ and $\varrho(V) = 0$ otherwise – in other words, that all newborn queues have the same initial volume V_0. To find the stationary solution in this case, we must solve Equation (5.2) when the left-hand side is equal to 0 and when $J(t)$ is equal to some constant J_0 that matches the exit flow. In this case, observe that the ansatz $f(V) = k + a^V$ is a solution, provided that

$$(\mu+\nu)a^2 - (\lambda+\mu+\nu)a + \lambda = 0. \quad (5.3)$$

Equation (5.3) has two roots,

$$a_+ = 1; \qquad a_- = \frac{\lambda}{\mu+\nu} \leq 1.$$

The solution that tends to zero at large V reads

$$P_{\text{st.}}(V \leq V_0) = A + Ba_-^V,$$
$$P_{\text{st.}}(V > V_0) = Ca_-^V,$$

with constants A, B, and C such that the following conditions hold:

(i) By considering $V = 1$, it must hold that

$$-(\lambda+\mu+\nu)P_{\text{st.}}(1) + (\mu+\nu)P_{\text{st.}}(2) = 0,$$

or

$$A\lambda = Ba_-[(\mu+\nu)a_- - (\lambda+\mu+\nu)],$$

which can be simplified further to $A = -B$.

(ii) By considering $V = V_0$, it must hold that

$$J_0 + (\mu+\nu)[(C-B)a_-^{V_0+1} - A] = 0.$$

(iii) By considering the exit flux, it must hold that

$$J_0 = (\mu+\nu)P_{\text{st.}}(1) = (\mu+\nu)(A + Ba_-).$$

(iv) By considering the normalisation of probabilities, it must hold that

$$\sum_{V>0} P_{\text{st.}}(V) = 1.$$

Rearranging these equations yields

$$C = A(a_-^{-V_0} - 1); \qquad J_0 = A(\mu+\nu-\lambda). \tag{5.4}$$

Finally, the normalisation condition simplifies to

$$AV_0 = 1. \tag{5.5}$$

The stationary solution $P_{\text{st.}}(V)$ then reads:

$$P_{\text{st.}}(V \leq V_0) = \frac{1}{V_0}\left(1 - a_-^V\right),$$
$$P_{\text{st.}}(V > V_0) = \frac{1}{V_0}\left(a_-^{-V_0} - 1\right)a_-^V.$$

The average volume $\bar{V} = \sum_{V=1}^{\infty} V P_{\text{st.}}(V)$ can be computed exactly for any V_0, but its expression is messy. It simplifies in the limit $a_- \to 1$, where it becomes:

$$\bar{V} \approx (1 - a_-)^{-1} = \frac{\mu + \nu}{\mu + \nu - \lambda}, \tag{5.6}$$

independently of V_0. This result clearly shows that the average volume diverges as $\lambda \uparrow (\mu + \nu)$, beyond which the problem admits no stationary state.

Beyond being a good exercise in dealing with master equations, the above calculations also provide insight into how we might approach the more general case, in which reinjection occurs not with some constant volume V_0, but instead with some arbitrary reinjection probability $\varrho(V_0)$. By linearity of the equations, we can solve this case as a linear superposition of the above solutions for different values of V_0, each with weight $\varrho(V_0)$. In this case, the normalisation in Equation (5.5) now reads:

$$A \sum_{V_0} V_0 \varrho(V_0) = 1. \tag{5.7}$$

How might we choose the function ϱ? One possible choice is simply to self-consistently use the stationary distribution, $P_{\text{st.}}(V)$. This corresponds to the simplifying assumption that whenever a queue depletes to 0, it is replaced by a new queue (behind it) drawn from the same stationary distribution. We will discuss this further in Section 5.3.6.

5.3.2 First-Hitting Times

Another quantity of interest is the length of time that elapses before a queue with a given length V first reaches $V = 0$. We call this time the **first-hitting time** $T_1(V)$. Given that we assume the queue dynamics to be stochastic, it follows that $T_1(V)$ is a random variable. In this section, we use the master equation (5.2) to derive several interesting results about its behaviour.

Let $\mathbb{E}[T_1|V]$ denote the mean first-hitting time from state V. We first turn our attention to the case where $V = V_0$. Recall that we can consider $V(t)$ as the state of a random walker with dynamics specified by Equation (5.1), and that the exit flux J_0 denotes the probability per unit time that the queue hits zero (and is reinjected at V_0). The quantities $\mathbb{E}[T_1|V_0]$ and J_0 are then related by the equation

$$J_0 = \frac{1}{\mathbb{E}[T_1|V_0]}. \tag{5.8}$$

To see why this is the case, imagine observing the system for a very long time T. By definition of the exit flux J_0, the expected number of hits during time T is equal to $J_0 T$. If the average length of time between two successive hits is $\mathbb{E}[T_1]$, then the average number of hits during T is also $T/\mathbb{E}[T_1]$. We thus obtain a simple

formula for the mean first-hitting time:

$$\mathbb{E}[T_1|V_0] = \frac{1}{J_0} = \frac{V_0}{\mu+\nu-\lambda}. \tag{5.9}$$

Intuitively, this result makes sense because $\mu+\nu-\lambda$ is the velocity with which the $V(t)$ process moves towards zero, and V_0 is the initial position.

We next turn our attention to deriving the full distribution of first-hitting times T_1. More precisely, we consider the question: given that the total volume in the queue is currently V, what is the probability $\Phi(\tau, V)$ that the first-hitting time is $T_1 = \tau$?

Let τ_1 denote the time at which the queue first changes volume (up or down). For the queue length to have remained constant between 0 and τ_1 requires that no event occurs during that time interval. Because of the Poissonian nature of volume changes, this happens with probability $e^{-(\nu+\mu+\lambda)\tau_1}$. Then, between times τ_1 and $\tau_1 + \mathrm{d}\tau_1$, the queue length will either increase with probability $\lambda \mathrm{d}\tau_1$ or decrease with probability $(\nu+\mu)\mathrm{d}\tau_1$. The first-hitting time problem then restarts afresh, but now from a volume $V+1$ or $V-1$ and first-hitting time $\tau - \tau_1$. Summing over these two different possibilities, one finds, for $V \geq 1$:

$$\Phi(\tau, V) = \int_0^\tau \mathrm{d}\tau_1 \lambda e^{-(\nu+\mu+\lambda)\tau_1} \Phi(\tau - \tau_1, V+1) \tag{5.10}$$

$$+ \int_0^\tau \mathrm{d}\tau_1 (\mu+\nu) e^{-(\nu+\mu+\lambda)\tau_1} \Phi(\tau - \tau_1, V-1), \tag{5.11}$$

with, trivially, $\Phi(\tau, V=0) = \delta(\tau)$.

By introducing the **Laplace transform** of $\Phi(\tau, V)$,

$$\widehat{\Phi}(z, V) := \int_0^\infty \mathrm{d}\tau \Phi(\tau, V) e^{-z\tau}, \tag{5.12}$$

and using standard manipulations (see Appendix A.2), one can transform this equation for $\Phi(\tau, V)$ to arrive at:

$$\widehat{\Phi}(z, V) = \frac{\lambda}{\lambda+\mu+\nu+z} \widehat{\Phi}(z, V+1) + \frac{\mu+\nu}{\lambda+\mu+\nu+z} \widehat{\Phi}(z, V-1). \tag{5.13}$$

One can check directly that $\widehat{\Phi}(z=0, V) = 1$ solves the above equation for $z = 0$. Recalling Equation (5.12), this just means that

$$\int_0^\infty \mathrm{d}\tau \Phi(\tau, V) = 1, \qquad \text{for all } V,$$

which implies that the probability that the queue will eventually deplete at some time in the future is equal to one (when $\lambda \leq \mu + \nu$).

By expanding Equation (5.12) for small z, we can also derive expressions for the moments of the distribution of first-hitting times:

$$\widehat{\Phi}(z,V) = 1 - z\mathbb{E}[T_1|V] + \frac{z^2}{2}\mathbb{E}[T_1^2|V] + O(z^3), \qquad (z \to 0). \tag{5.14}$$

Expanding Equation (5.13) to first-order in z therefore leads to:

$$-z\mathbb{E}[T_1|V] = -z\frac{\lambda}{\lambda+\mu+\nu}\mathbb{E}[T_1|V+1] - z\frac{\mu+\nu}{\lambda+\mu+\nu}\mathbb{E}[T_1|V-1] - z\frac{1}{\lambda+\mu+\nu}, \tag{5.15}$$

which implies that:

$$\mathbb{E}[T_1|V] = \frac{\lambda}{\lambda+\mu+\nu}\mathbb{E}[T_1|V+1] + \frac{\mu+\nu}{\lambda+\mu+\nu}\mathbb{E}[T_1|V-1] + \frac{1}{\lambda+\mu+\nu}. \tag{5.16}$$

This has a transparent interpretation: the average time for the queue to reach length 0 from an initial length of V is equal to the average time to make a move up or down, plus the average time to reach 0 from $V+1$ times the probability of making an upward move, plus the average time to reach 0 from $V-1$ times the probability of making a downward move.

In the case where $\lambda < \mu + \nu$, we can interpret the difference between $(\mu+\nu)$ and λ as the mean drift of the queue length towards 0. Since this drift does not depend on V, this suggests that when seeking solutions to $\mathbb{E}[T_1|V]$, we should consider expressions that take the form

$$\mathbb{E}[T_1|V] = A'V, \tag{5.17}$$

where, using Equation (5.16),

$$A'(\lambda - \mu - \nu) + 1 = 0. \tag{5.18}$$

Substituting Equation (5.18) into Equation (5.17), we finally recover

$$\mathbb{E}[T_1|V] = \frac{V}{\mu+\nu-\lambda},$$

which is exactly what we derived in Equation (5.9) by a different method.

An interesting point is that Equation (5.13) can be solved for any z, and that $\widehat{\Phi}(z,V) = a(z)^V$ is a solution, provided that

$$\lambda a^2(z) - (\lambda+\mu+\nu+z)a(z) + \mu + \nu = 0, \tag{5.19}$$

which can be rearranged to

$$a_\pm(z) = \frac{1}{2\lambda}\left[(\lambda+\mu+\nu+z) \pm \sqrt{(\lambda+\mu+\nu+z)^2 - 4\lambda(\mu+\nu)}\right]. \tag{5.20}$$

Note that $a_+(0) = (\mu+\nu)/\lambda > 1$ and $a_-(0) = 1$. Hence, the constraint $\widehat{\Phi}(0,V) \equiv 1$ eliminates the exponentially growing contribution $a_+(z)^V$, which finally leads to a very simple result:

$$\ln \widehat{\Phi}(z,V) = V \ln a_-(z). \tag{5.21}$$

Expanding to second order in z directly allows one to derive the variance of the first-hitting time (see Appendix A.2):

$$\mathbb{V}[T_1|V] = V \frac{\mu+\nu+\lambda}{(\mu+\nu-\lambda)^3}. \tag{5.22}$$

If an explicit solution is required, Equation (5.21) can be inverted to yield the exact hitting time distribution for any V. However, the analytical form of this solution is messy and not particularly instructive. More interestingly, Equation (5.21) shows that all cumulants of the hitting time distribution behave linearly in V. This situation is identical to the well-known **Central Limit Theorem**, where all cumulants of the sum of V IID random variables behave linearly in V (see Appendix A.2).[1] Therefore, Equation (5.21) can be used to show that for large V, the distribution of first-hitting times approaches a Gaussian distribution with mean $V/(\mu+\nu-\lambda)$ and variance $V(\mu+\nu+\lambda)/(\mu+\nu-\lambda)^3$.

5.3.3 Long Lifetimes: The Critical Case

So far in this section, we have considered the case where $\lambda < \mu+\nu$. The case $\lambda = \mu+\nu$, where the flow is exactly balanced, is quite different. By inspection, one sees that the expression for the mean first-hitting time in Equation (5.9) diverges when $\lambda = \mu+\nu$. Therefore, the expected first-hitting time in this case is infinite. However, this does not imply that the queue never empties. In fact, the queue empties with probability 1, although this can take a very long time.

In this special case, expanding Equations (5.20) and (5.21) in the limit $z \to 0$ leads to the singular expansion:

$$\ln \widehat{\Phi}(z,V) = V \ln a_-(z) \underset{z\to 0}{\approx} -V\sqrt{\frac{z}{\lambda}} + O(z). \tag{5.23}$$

The $\sqrt{z}$ behaviour for small z implies that for large τ, the distribution of the hitting time $\Phi(\tau,V)$ decays as $(V/\sqrt{\lambda})\tau^{-3/2}$ (see Appendix A.2). This distribution is very broad; in fact, it has an infinite mean, which is characteristic of unbiased one-dimensional random walks. This behaviour can be traced back to the very long upward excursions of the queue length, as the force driving to make it smaller is exactly zero in the critical case. When $\lambda \uparrow (\mu+\nu)$, the $\tau^{-3/2}$ tail is truncated beyond a time scale $\propto (\mu+\nu-\lambda)^{-2}$, much larger than the average exit time $\mathbb{E}[T_1|V]$.

[1] In fact, the first-hitting time from V is equal to the first-hitting time of $V-1$ from V plus the first-hitting time from $V-1$, which recursively demonstrates that $T_1(V)$ is indeed the sum of V independent random variables.

In fact, the only way that the above (highly simplified) random walk model can be used to represent the dynamics of long queues is when the parameters are chosen to be close to the critical case – otherwise, order queues would never become large – see Equation (5.6). However, as we will see in the subsequent sections of this chapter, this fine-tuning to a critical case is not necessary when we consider more complex models in which the aggregate cancellation rate increases with the total volume of limit orders in the queue.

5.3.4 The Continuum Limit and the Fokker–Planck Equation

Throughout this section, we have studied a model in which the state space of V is discrete, and where V can only change by one unit at a time. We now consider the limit where the volume in the queue becomes large, such that $V \gg 1$ most of the time. In this situation, volume changes can be considered as infinitesimal. In this **continuum limit** (where V is now considered as a continuous variable), the master equation from Section 5.3.1 instead becomes a **Fokker–Planck equation**.

Consider again the dynamics described by Equation (5.1), and assume that $P(V,t)$ is sufficiently smooth to allow the following Taylor expansion:

$$P(V \pm 1, t) \approx P(V,t) \pm \partial_V P(V,t) + \frac{1}{2}\partial^2_{VV} P(V,t). \tag{5.24}$$

Substituting this expansion into the master equation (5.2) with $J(t) = 0$:

$$\begin{aligned}\frac{\partial P(V,t)}{\partial t} &\approx -(\lambda+\mu+\nu)P(V,t) + \lambda\left(P(V,t) - \partial_V P(V,t) + \frac{1}{2}\partial^2_{VV}P(V,t)\right) \\ &\quad + (\mu+\nu)\left(P(V,t) + \partial_V P(V,t) + \frac{1}{2}\partial^2_{VV}P(V,t)\right), \\ &\approx -(\lambda-\mu-\nu)\partial_V P(V,t) + (\lambda+\mu+\nu)\frac{1}{2}\partial^2_{VV}P(V,t).\end{aligned}$$

Introducing the **drift constant** $F = \lambda - \mu - \nu$ and the **diffusion constant** $D = (\lambda + \mu + \nu)/2$ yields

$$\frac{\partial P(V,t)}{\partial t} \approx -F\partial_V P(V,t) + D\partial^2_{VV}P(V,t). \tag{5.25}$$

Equation (5.25) is the simplest case of a Fokker–Planck equation, called the **drift–diffusion equation**. Working directly with this equation provides an alternative (and often simpler) route to calculating some of the quantities and distributions that we calculated for the case of strictly positive order sizes earlier in this section. For example, let us consider the stationary distribution $P_{\text{st.}}(V)$, which we calculate in Section 5.4.1 by solving the discrete master equation. We now consider Equation (5.25) with the left-hand side set to 0. We add an extra current contribution $J_0\delta(V - V_0)$ to the right-hand side, to account for the reinjection of

V_0 when the queue depletes to 0. For $F < 0$, the solution such that $P_{\text{st.}}(V = 0) = 0$ (because of the absorbing condition there) and $P_{\text{st.}}(V \to \infty) = 0$ takes the form:

$$P_{\text{st.}}(V \leq V_0) = A(1 - e^{-|F|V/D}),$$
$$P_{\text{st.}}(V \geq V_0) = A(e^{|F|V_0/D} - 1)e^{-|F|V/D}.$$

The discontinuity of the slope at $V = V_0$ is fixed by the extra current contribution $J_0\delta(V - V_0)$:

$$A|F|e^{-|F|V_0/D} - J_0 = -A|F|(1 - e^{-|F|V_0/D}),$$

which yields

$$A = J_0/|F|.$$

Finally, the normalisation of $P_{\text{st.}}(V)$ fixes A such as $AV_0 = 1$, leading to $J_0 = |F|/V_0$. The mean first-hitting time is thus given by $\mathbb{E}[T_1] = V_0/|F|$.

5.3.5 Revisiting the Laplace Transform

We can also consider the continuum limit within the Laplace transform in Equation (5.12), to find:

$$s\widehat{\Phi}(z, V) = F\partial_V\widehat{\Phi}(z, V) + D\partial^2_{VV}\widehat{\Phi}(z, V). \tag{5.26}$$

Looking for solutions of the form $\widehat{\Phi}(z, V) \propto e^{a(z)V}$, it holds that $a(z)$ must obey

$$Da^2(z) + Fa(z) - z = 0,$$

so

$$a_{\pm}(z) = \frac{1}{2D}\left[-F \pm \sqrt{F^2 + 4Dz}\right].$$

Note that the problem is only well defined when $F < 0$ (i.e. when the drift is towards $V = 0$). The constraint that $\widehat{\Phi}(z = 0, V) = 1$ for all V eliminates the positive $a_+(z)$ solution. Hence

$$\ln\widehat{\Phi}(z, V) = a_-(z)V, \tag{5.27}$$

which is very similar to Equation (5.21). Note that $a_-(0) = 0$, such that the condition $\ln\widehat{\Phi}(0, V) = 0$ is indeed satisfied.

5.3.6 A Self-Consistent Solution

We now discuss the *self-consistent* case, in which reinjection occurs according to the stationary distribution $P_{\text{st.}}(V)$, such that the probability of observing a reinjection of a given size is precisely equal to the stationary probability of observing a queue of that size. This corresponds to the idea that when a queue empties, it is replaced by the queue just behind, which we assume to have the same stationary distribution. This is of course only an approximation, since the second-best queue is shielded from

market orders, and might not be described by the same deposition and cancellation rates as the best queue.

The Fokker–Planck equation for $P_{\text{st.}}(V)$ reads

$$\frac{\partial P_{\text{st.}}(V)}{\partial t} = -F\partial_V P_{\text{st.}}(V) + D\partial^2_{VV} P_{\text{st.}}(V) + J_0 P_{\text{st.}}(V),$$

and by stationarity of $P_{\text{st.}}(V)$ it holds that

$$\frac{\partial P_{\text{st.}}(V)}{\partial t} = 0,$$

so we seek to solve

$$-F\partial_V P_{\text{st.}}(V) + D\partial^2_{VV} P_{\text{st.}}(V) + J_0 P_{\text{st.}}(V) = 0. \tag{5.28}$$

To find the solution that vanishes at $V = 0$, we proceed exactly as in the previous section, but replacing $\widehat{\Phi}(z, V)$ in Equation (5.26) with $P_{\text{st.}}$, replacing z with $-J_0$, and replacing F with $-F$. Since now both solutions $a_+(-J_0)$ and $a_-(-J_0)$ are negative for $F < 0$, we have:

$$P_{\text{st.}}(V) = A(-e^{a_+(-J_0)V} + e^{a_-(-J_0)V}).$$

The exit flux at $V = 0$ is given by $J_0 = -D\partial_V P_{\text{st.}}$, leading to:

$$J_0 = DA(a_+(-J_0) - a_-(-J_0)) = A\sqrt{F^2 - 4DJ_0}.$$

The normalisation condition is

$$1 = -A\left(\frac{1}{a_+(-J_0)} - \frac{1}{a_-(-J_0)}\right),$$

which turns out to be satisfied for any A. We can thus arbitrarily choose A, which amounts to choosing an arbitrary exit flux $J_0 \leq F^2/4D$. The corresponding stationary solution reads:

$$P_{\text{st.}}(V) = \frac{2J_0}{\sqrt{F^2 - 4DJ_0}} e^{-|F|V/2D} \sinh\left(\frac{\sqrt{F^2 - 4DJ_0}\, V}{2D}\right). \tag{5.29}$$

Interestingly, this model can self-consistently produce both long queues and short queues in the limit $J_0 \to 0$.

5.4 A More Complex Model: Linear Cancellation Rate

Throughout Section 5.3, we assumed that the total rate of cancellations was constant. Although this framework led to convenient mathematics, it is not a realistic representation of cancellations in real markets. In particular, the model requires fine-tuning of its parameters to produce long queues, such as those often observed in empirical data for large-tick assets.

How might we change the model to address this problem? One way is to assume that each individual order in the queue has cancellation rate ν independent of the queue length V, such that the total cancellation rate grows linearly as νV (instead of being constant, as in the previous model). This makes intuitive sense: the more

orders in the queue, the larger the probability that one of them gets cancelled. Specifically, consider the model

$$\begin{cases} V \to V+1 & \text{with rate } \lambda \mathrm{d}t \quad \text{(deposition)}, \\ V \to V-1 & \text{with rate } (\mu + V\nu)\mathrm{d}t \quad \text{(execution + cancellation)}. \end{cases} \tag{5.30}$$

To lighten exposition in the subsequent sections, we introduce the notation $W_{\pm}(V)$ as the **transition rate** from $V \to V \pm 1$. For the model described by Equation (5.30),

$$W_{+}(V) = \lambda, \tag{5.31}$$

$$W_{-}(V) = \mu + V\nu. \tag{5.32}$$

Observe that the only difference between this model and the model specified by Equation (5.1) is the inclusion of the $V\nu$ term, instead of the ν term, to account for cancellations.

Importantly, the inclusion of this $V\nu$ term ensures that queue sizes self-stabilise for any choices of $\lambda, \mu, \nu > 0$. In other words, there exists some value

$$V^* = \frac{\lambda - \mu}{\nu} \tag{5.33}$$

of the queue length such that the total out-flow rate of orders $W_{-}(V^*)$ matches the in-flow rate of orders $W_{+}(V^*)$. If $V < V^*$, then $W_{+}(V) > W_{-}(V)$ and the queue length tends to grow on average; if $V > V^*$, then $W_{+}(V) < W_{-}(V)$ and the queue length tends to shrink on average. Those self-regulating mechanisms lead to queues that fluctuate around their **equilibrium size** V^*. Given a large value of λ, it is therefore possible to observe long queues without requiring a fine-tuned relation between λ, μ, and ν to ensure stability (as was the case in Section 5.3).

5.4.1 Quasi-Equilibrium

The master equation corresponding to the dynamics specified by Equation (5.30) is now given by

$$\frac{\partial P(V,t)}{\partial t} = -(\lambda + \mu + \nu V)P(V,t) + \lambda P(V-1,t) + (\mu + \nu(V+1))P(V+1,t); \quad V > 1. \tag{5.34}$$

For the moment, we fix the boundary condition such that whenever $V = 0$, the particle is immediately reinjected at state $V = 1$. Therefore, the exit flux is 0, there is no particle at $V = 0$ and the missing equation for $V = 1$ reads

$$\frac{\partial P(1,t)}{\partial t} = -\lambda P(1,t) + (\mu + 2\nu)P(2,t).$$

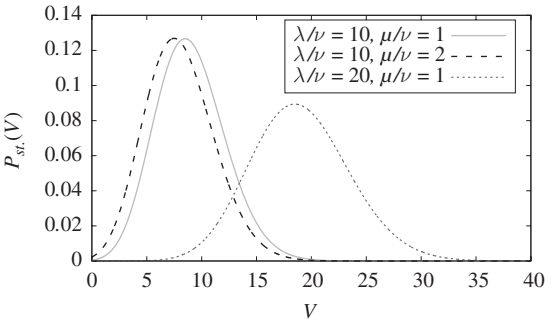

Figure 5.1. $P_{\text{st.}}(V)$ from Equation (5.35) for different values of λ/ν and μ/ν.

In this case, one can readily check that for $V \geq 2$, the ansatz

$$P_{\text{st.}}(V) = A \frac{\left(\frac{\lambda}{\nu}\right)^V}{\Gamma\left[V + 1 + \frac{\mu}{\nu}\right]} \tag{5.35}$$

is an exact stationary state, in the sense that it causes the right-hand side of Equation (5.34) to equal 0 for all $V > 1$. The value of $P_{\text{st.}}(1)$ is simply given by $(\mu + 2\nu)P_{\text{st.}}(2)$ and the constant A is fixed by the normalisation condition $\sum_{V=1}^{\infty} P_{\text{st.}}(V) = 1$.

To make sense of this solution when depletion and reinjection are present, let us assume that we are in a situation where queues are typically very large, corresponding to the limit $\lambda \gg \mu, \nu$, such that $V^* \gg 1$. Figure 5.1 shows the shape of $P_{\text{st.}}(V)$ for this case. As the figure illustrates, the stationary distribution is peaked, with a very low probability of small queues.

More precisely, consider the relative probability of observing the queue in state $V = 1$ rather than in state $V = V^*$. This is given by

$$\frac{P_{\text{st.}}(1)}{P_{\text{st.}}(V^*)} \propto V^{*\frac{1}{2}+\frac{\mu}{\nu}} e^{-V^*}. \tag{5.36}$$

For increasingly large choices of V^*, this ratio becomes very small, very quickly. This suggests that even if we apply a boundary condition with non-zero exit flux, the contribution of this exit flux (which is proportional to $P_{\text{st.}}(V = 1)$) to the overall dynamics will be very small, so the stationary state will still be given by Equation (5.35) up to exponentially small corrections $\propto e^{-V^*}$.

As we discuss in Section 5.4.5, the order of magnitude of the time required to reach the stationary distribution is ν^{-1}. Equation (5.36) suggests that compared to ν^{-1}, the first-hitting time $T_1(V)$ is very large – indeed, it is exponentially larger in V^*, as we will confirm in Equation (5.42) below. Therefore, even in the presence of a non-zero exit flux, the system reaches a **quasi-equilibrium** (where

the distribution is approximately given by $P_{\text{st.}}(V)$) much faster than the time needed for the queue to empty.

5.4.2 First-Hitting Times

We now study the first-hitting time for the model with a constant cancellation rate per order. Using the notation $W_+(V)$ and $W_-(V)$ from Equations (5.31) and (5.32), we can implement the Laplace transform from Equation (5.12) to arrive at

$$\widehat{\Phi}(z,V) = \frac{W_+(V)}{W_+(V)+W_-(V)+z}\widehat{\Phi}(z,V+1) + \frac{W_-(V)}{W_+(V)+W_-(V)+z}\widehat{\Phi}(z,V-1). \quad (5.37)$$

As we did for the constant-cancellation model in Section 5.3.2, we can expand Equation (5.37) to first order in z to derive an expression for the mean first-hitting time $\mathbb{E}[T_1|V]$:

$$(W_+(V)+W_-(V))\mathbb{E}[T_1|V] - W_+(V)\mathbb{E}[T_1|V+1] - W_-(V)\mathbb{E}[T_1|V-1] = 1. \quad (5.38)$$

Setting

$$U(V) := \nu[\mathbb{E}[T_1|V+1] - \mathbb{E}[T_1|V]]$$

yields the recursion

$$W_+(V)U(V) - W_-(V)U(V-1) = -\nu. \quad (5.39)$$

Note that from the definition of $U(V)$, the mean first-hitting time of state V when starting from state $V+1$ is $U(V)/\nu$.

The recursion in Equation (5.39) can be solved by using the generating function

$$\widetilde{U}(y) = \sum_{V=0}^{\infty} \frac{y^V}{V!}U(V), \quad (5.40)$$

which obeys a simple ordinary linear differential equation.[2] The problem then resides in finding appropriate boundary conditions to impose on the solution. When $y \to 0$, the solution should be 0, but when $y \to \infty$, the solution should not grow exponentially in y.

In general, finding a solution that satisfies these boundary conditions is difficult. However, when $\mu/\nu \ll 1$, the solution is simple. As can be verified by inserting it into Equation (5.39), the solution reads:[3]

$$U(V) = \frac{V!}{V^{*V+1}} \sum_{K=V+1}^{\infty} \frac{V^{*K}}{K!}. \quad (5.41)$$

Equation (5.41) contains all the information we need to infer how the mean first-hitting time depends on the volume of the queue. When $V^* \gg 1$ (i.e. when the mean queue length is large), the function $U(V)$ has two regimes:

$$U(V < V^*) \approx \frac{V!}{V^{*V+1}}e^{V^*},$$

$$U(V > V^*) \approx \frac{1}{V}.$$

[2] Another possible method for solving Equation (5.39) is to use continued fractions (see, e.g., Cont, R., Stoikov, S., & Talreja, R. (2010). A stochastic model for order book dynamics. *Operations Research*, 58(3), 549–563). For a related calculation, see also Godrèche, C., Bouchaud, J. P., & Mézard, M. (1995). Entropy barriers and slow relaxation in some random walk models. *Journal of Physics A*, 28, L603–L611.

[3] For arbitrary μ/ν, the solution is obtained by replacing $V!$ by $\Gamma[V+1+\mu/\nu]$ and $K!$ by $\Gamma[K+1+\mu/\nu]$. This does not change the qualitative behaviour of $\mathbb{E}[T_1|V]$ when $V^* \gg 1$.

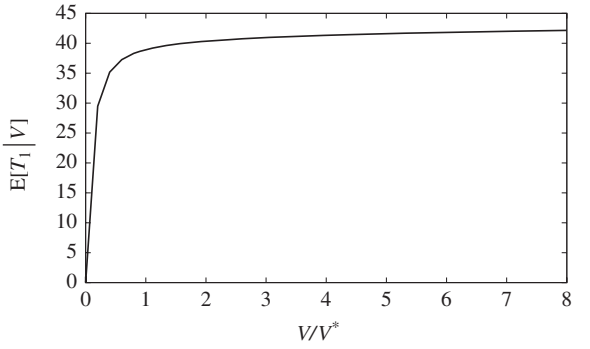

Figure 5.2. The average first-hitting time $\mathbb{E}[T_1|V]$ versus V/V^* for $V^* = 5$. A quasi-plateau value $\approx e^{V^*}/\nu V^*$ is reached quickly as the initial volume V increases.

Observing that

$$\nu\mathbb{E}[T_1|V] = U(0) + U(1) + \cdots + U(V-1)$$

allows us to infer the behaviour of the mean first-hitting time as a function of V. As Figure 5.2 illustrates, $\mathbb{E}[T_1|V]$ grows rapidly towards a quasi-plateau that extends over a very wide region around V^*, before finally growing as $\ln V$ when $V \to \infty$. We provide an intuitive explanation for this logarithmic behaviour in Section 5.4.4, where we discuss the continuous limit of the model.

The plateau value defines the mean first-hitting time when starting from a queue size $V_0 \approx V^*$, and is given by

$$\mathbb{E}[T_1|V^*] \approx \frac{e^{V^*}}{\nu V^*} \times \left[1 + \frac{1!}{V^*} + \frac{2!}{(V^*)^2} + \cdots\right]. \tag{5.42}$$

The mean first-hitting time thus grows exponentially in V^*, as anticipated by the arguments of Section 5.4.1.

5.4.3 Poissonian Queue Depletions

We now show that in the limit when $\mathbb{E}[T_1|V^*]$ is large, the distribution of first-hitting times is exponential, such that the counting process of queue depletions is Poissonian. Consider again Equation (5.37) for the Laplace transform of the distribution of first-hitting times. Expanding $\widehat{\Phi}(z, V)$ in powers of z generates the moments of this distribution (see Appendix A.2):

$$\widehat{\Phi}(z, V) = 1 + \sum_{n=1}^{\infty} \frac{(-z)^n}{n!}\mathbb{E}[T_1^n|V]. \tag{5.43}$$

Identifying terms of order z^n in Equation (5.37) leads to the following exact recursion equation:

$$\begin{aligned}&(W_+(V) + W_-(V))\mathbb{E}[T_1^n|V] - W_+(V)\mathbb{E}[T_1^n|V+1] - W_-(V)\mathbb{E}[T_1^n|V-1] = \\ &n! \sum_{k=1}^{n} \frac{(W_+(V) + W_-(V))^{-k}}{(n-k)!} \left[W_+(V)\mathbb{E}[T_1^{n-k}|V+1] + W_-(V)\mathbb{E}[T_1^{n-k}|V-1]\right].\end{aligned} \tag{5.44}$$

When $n = 1$, we recover Equation (5.38). When $n = 2$, and for V in the wide plateau region (where the average hitting time is nearly constant) defined in Section 5.4.2, the right-hand side of Equation (5.44) contains two types of terms:

- $k = 1$: terms of order $\mathbb{E}[T_1|V^*]$, which grow exponentially in V^*;
- $k = 2$: terms of order $\mathbb{E}[T_1^0|V]$, which are of order unity.

Neglecting the latter contribution, we find

$$(W_+(V) + W_-(V))\mathbb{E}[T_1^2|V] - W_+(V)\mathbb{E}[T_1^2|V+1] - W_-(V)\mathbb{E}[T_1^2|V-1] \approx 2\mathbb{E}[T_1|V^*].$$

Therefore the equation for $\mathbb{E}[T_1^2|V]$ is exactly the same as for $\mathbb{E}[T_1|V]$, except that the right-hand side is now equal to $2\mathbb{E}[T_1|V^*]$ instead of 1. Since the equation is linear, we can conclude that

$$\mathbb{E}[T_1^2|V] \approx 2\mathbb{E}[T_1|V^*]\mathbb{E}[T_1|V],$$

which again does not depend heavily on V in a wide region around V^*. We can then extend this argument by recursion to arbitrary n. Up to the leading exponential contribution, the dominant term is always $k = 1$, hence

$$\mathbb{E}[T_1^n|V] \approx n!\mathbb{E}[T_1|V]^n,$$

which are the moments of an exponential distribution with mean $\mathbb{E}[T_1|V]$. Consequently, the distribution of hitting times is

$$\Phi(\tau, V) \approx \frac{1}{\mathbb{E}[T_1|V]} \exp -\left[\frac{\tau}{\mathbb{E}[T_1|V]}\right].$$

Consistently with the physical picture that we develop in Section 5.4.5, this result illustrates that the sequences of events that cause the queue to empty occur so rarely that we can regard them as independent, leading to a Poisson depletion process. This rather remarkable property will enable us to discuss the race between two queues in a very simple manner in Chapter 7.

5.4.4 The Continuum Limit: The CIR Process

We now consider the continuum limit for the model with a constant cancellation rate per order. Similarly to our approach in Section 5.3.4, we assume that $P(V,t)$ is sufficiently smooth to allow the same Taylor expansion as in Equation (5.24). The Fokker–Planck equation describing the linear problem is then:

$$\frac{\partial P(V,t)}{\partial t} \approx -\partial_V [F(V)P(V,t)] + \partial_{VV}^2 [D(V)P(V,t)], \tag{5.45}$$

with drift

$$F(V) = W_+(V) - W_-(V) = \lambda - \mu - \nu V$$

and diffusion

$$D(V) = \frac{W_+(V) + W_-(V)}{2} = \frac{\lambda + \mu + \nu V}{2}.$$

The equilibrium size,

$$V^* = (\lambda - \mu)/\nu, \tag{5.46}$$

corresponds to the point where the drift $F(V)$ vanishes.

This model is known as a **modified CIR process**.[4] Given that we are considering the specific context of a queue, we will call the model the Q-CIR process (where

[4] The model is also sometimes called a Heston process or a mean-reverting Bessel process. The associated stochastic PDE, defining the dynamics of V, is $\mathrm{d}V = (\lambda - \mu - \nu V)\mathrm{d}t + \sqrt{\lambda + \mu + \nu V}\mathrm{d}W_t$, where W_t is a Brownian motion. The standard CIR process is obtained upon defining $X := \lambda + \mu + \nu V$.

Q stands for queuing). Note that the model holds quite generally; for example, it is not necessary to assume that all orders have the same size to obtain such a Q-CIR process in the continuum limit.

If we impose an absorbing boundary condition at $V = 0$, then, in the absence of reinjection, Equation (5.45) does not have any non-trivial stationary solutions, because for all $V > 0$, it follows that $\lim_{t\to\infty} P(V,t) = 0$.

If we instead impose a reflecting boundary condition at $V = 0$, such that queues are not allowed to empty and instead bounce back to a strictly positive value, then the exit flux at $V = 0$ is zero, so the stationary solution of Equation (5.45) obeys

$$-F(V)P_{\text{st.}}(V) + \partial_V \left[D(V)P_{\text{st.}}(V)\right] = 0.$$

Solving this expression, we arrive at

$$P_{\text{st.}}(V) = A\left(\frac{\lambda+\mu}{\nu} + V\right)^{\frac{4\lambda}{\nu}-1} e^{-2V},$$

where A is fixed by the normalisation of $P_{\text{st.}}(V)$.

In the case where the deposition rate λ becomes very large, such that $\mu \ll \lambda$ and $V^* \approx \lambda/\nu \gg 1$, the stationary state can be written as

$$P_{\text{st.}}(V) \approx A\,(V^* + V)^{4V^*} e^{-2V},$$

which behaves very similarly to its discrete counterpart in Equation (5.35) (see also Figure 5.1). In this case, $P_{\text{st.}}(V)$ first grows exponentially (as e^{2V} for $V \ll V^*$), before reaching a maximum at $V = V^*$ and decaying back to zero as $V^{4V^*}e^{-2V}$ for $V \gg V^*$. Similarly to in Equation (5.36), we can calculate the following ratio of probabilities

$$\frac{P_{\text{st.}}(0)}{P_{\text{st.}}(V^*)} \approx e^{-(4\ln 2-2)V^*}.$$

As in the discrete model, this ratio behaves like e^{V^*}, so for increasingly large choices of V^*, this ratio becomes very small, very quickly. However, note that the continuum limit predicts a coefficient of $4\ln 2 - 2 \approx 0.76$ in the exponential, whereas the discrete case in Equation (5.36) predicts a coefficient of 1 in the exponential. This highlights that adopting the assumptions required to study the model in the continuum limit can induce subtle quantitative differences that may affect the evaluation of rare events.

Despite these quantitative differences, the continuum limit allows us to gain useful intuition about the dynamics of the process. The reason for the exponentially small probability of observing small queues is quite simple: since the drift $F(V)$ is strongly positive when $V \ll V^*$, it tends to drive the queue back to its equilibrium size V^*. In the same way, when $V \gg V^*$, the drift $F(V)$ is negative, and tends to

drive the queue back to V^*. Correspondingly, it takes a very long time to overcome this drift and reach $V = 0$.

In the limit $\lambda \gg \mu, \nu$, one can in fact show that

$$\mathbb{E}[T_1|V] \approx \nu^{-1} \sqrt{\frac{\pi}{2V^*}}\, e^{(4\ln 2 - 2)V^*}, \tag{5.47}$$

which again shows very little dependence on V in a wide region around V^*.

Similarly to our findings in Section 5.4.1, we recover the result that in the presence of a non-zero exit flux, the system reaches a quasi-equilibrium (where the distribution is approximately given by $P_{\text{st.}}(V)$) exponentially faster than the time needed for the queue to empty. In fact, this result still holds when re-initialising the queue size to almost any value V_0 after the queue depletes to $V = 0$, because the queue size re-equilibrates to $P_{\text{st.}}(V)$ so quickly (in a time whose magnitude is of the order ν^{-1}). This result is important because it illustrates that it is not necessary to specify the exact reinjection mechanism to obtain a precise prediction for the shape of the stationary distribution $P_{\text{st.}}(V)$. In practice, the distribution of initial values V_0 should be the size distribution of either the second-best queue, or of an incipient new queue (depending on the cases; see Section 6.1 below). But in both cases, $P_{\text{st.}}(V)$ will quickly settle in, with very little dependence on the re-initialisation mechanism.

5.4.5 *Effective Potential Barriers*

How can one estimate the mean first-hitting time using the Fokker–Planck formalism? In the continuum limit, the analogues of Equations (5.37) and (5.38) read

$$z\widehat{\Phi}(z, V) = F(V)\partial_V \widehat{\Phi}(z, V) + D(V)\partial_{VV}^2 \widehat{\Phi}(z, V) \tag{5.48}$$

and

$$F(V)\partial_V \mathbb{E}[T_1|V] + D(V)\partial_{VV}^2 \mathbb{E}[T_1|V] = -1, \tag{5.49}$$

respectively, with boundary conditions

$$\mathbb{E}[T_1|V = 0] = 0; \qquad \lim_{V\to\infty} \partial_V \mathbb{E}[T_1|V] = 0.$$

This second-order, inhomogeneous ODE can be solved by standard techniques, to produce the result shown in Equation (5.47).

As in the discrete case, the mean first-hitting time is only weakly dependent on the starting size V_0 (provided V_0 is neither too large or too small) and grows exponentially in the equilibrium size V^*. To understand these (perhaps surprising) results, we provide the following physical interpretation of the system. Consider a Brownian particle (with a V-dependent diffusion constant) in a parabolic-like potential well that reaches its minimum at V^* with curvature $\mathcal{W}''(V^*) = 1/V^*$ (see

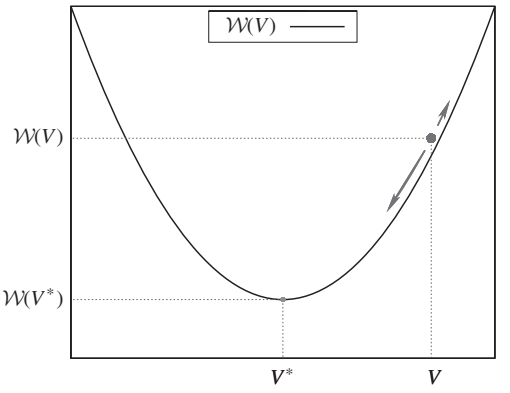

Figure 5.3. A potential well $\mathcal{W}(V)$, within which the effective "particle" (which represents the size of the queue) oscillates randomly before overcoming the barrier.

Figure 5.3). Consider the particle's **effective potential energy** $\mathcal{W}(V)$, defined such that $F(V) := -D(V)\mathcal{W}'$, and with $\mathcal{W}(0) = 0$. In the present linear case, and again in the limit $\lambda \gg \mu$, it follows that

$$\mathcal{W}(V) = 2V - 4V^* \ln\left(1 + \frac{V}{V^*}\right). \tag{5.50}$$

Intuitively, the particle's typical trajectories are such that starting from almost any V_0, the particle first rolls down the potential well, to reach the minimum after a time whose magnitude is of order ν^{-1}. Then, driven by the Brownian noise, the particle fluctuates around this minimum. It takes an extremely rare fluctuation to allow the particle to climb all the way uphill (from V^* to $V = 0$) and find the exit. As we saw in Section 5.4.3, these fluctuations in fact lead to a Poissonian exit process. The mean escape time of the particle is given by

$$\mathbb{E}[T_1|V] \approx \frac{1}{D(V^*)|\mathcal{W}'(0)|}\sqrt{\frac{2\pi}{\mathcal{W}''(V^*)}}e^{-\mathcal{W}(V^*)}.$$

In the Q-CIR case (see Section 5.4.4), where $\mathcal{W}$ is given by Equation (5.50), one can check that this expression recovers exactly Equation (5.47).

Finally, our physical picture can also shed light on the logarithmic behaviour of $\mathbb{E}[T_1|V]$ for $V \to \infty$. In this case, the Brownian particle must first scurry from this large V down to V^*, from which it then starts attempting to escape the potential well. The initial roll-down phase is governed by a quasi-deterministic evolution

$$\left.\frac{dV}{dt}\right|_{V\to\infty} \approx F(V) \approx -\nu V.$$

In this regime, we can neglect the Brownian noise contribution, because it is so small compared to the drift. Integrating this equation of motion reveals why it

takes a time whose magnitude is of order $\ln V$ to reach the bottom of the potential well, hence explaining the logarithmic growth of $\mathbb{E}[T_1|V]$ for large V.

5.5 Conclusion

In this chapter, we have considered many useful ideas about single-queue systems. Many of these results hold more generally, independently of the details of the model used to describe the dynamics of long queues. To summarise the most important intuitions that we have garnered:

- When the total cancellation rate grows with the queue size V, queues self-stabilise around a well-defined equilibrium size V^*.
- When V^* is sufficiently large (i.e. when the order submission process is more intense than either cancellations or market order execution), the queue dynamics can be neatly decomposed into a fast part, where a quasi-equilibrium state quickly sets in, and a slow part, where a Poisson process leads to an emptying of the queue.
- Although we have only considered the case where the cancellation rate grows linearly with the queue size, the resulting separation of time scales is in fact generic and holds whenever the total cancellation rate grows with the queue size, say as V^ζ with $\zeta > 0$.
- The case $\zeta = 0$ corresponds to the first model that we considered in this chapter. It is special in the sense that the only way to obtain long queues is to fine-tune the parameters of the model such that it sits at its "critical point", beyond which no stationary solution exists.

Empirical data suggests that the total cancellation rate grows *sublinearly* with the queue size (i.e. $0 < \zeta < 1$; see Section 6.4). A simple argument to understand this is time priority, which is extremely valuable in large-tick markets, because being at the front of a long queue increases the probability of execution, while minimising adverse selection (see Section 21.4 for a detailed discussion of this point). This means that cancelling an order at the front of a long queue does not make much sense economically. The only orders likely to be cancelled are those at the back of the queue. Therefore, we expect that only a small fraction of the orders in a given queue have an appreciable probability of being cancelled.

In the next chapter, we will consider more general specifications of single-queue models, and discuss how to calibrate such models directly to empirical data, without making many assumptions on the microscopic order flow. The above conclusions will hold for a large family of such models provided the effective potential barrier introduced in Section 5.4.5 is sufficiently high.

Take-Home Messages

(i) Stochastic models can be used to model markets with a "zero-intelligence" framework. These models seek to make predictions based on the statistical properties of order flow, rather than trying to understand what drives it in the first place.

(ii) The volume available at each level of the price grid can be modelled as a queuing system. These queues grow due to the arrival of new limit orders and shrink due to the cancellation of existing limit orders. At the best quotes, arriving market orders also cause the queues to shrink.

(iii) When all are described by mutually independent, homogeneous Poisson processes whose rate parameters are independent of the queue size, many quantities of interest can be derived analytically. Only when the total growth rate is nearly equal to the total shrink rate does the queue exhibit large excursions, as observed empirically. In this case, the distribution of times between two queue depletions is extremely broad.

(iv) When assuming a constant cancellation rate *per limit order*, the queue stabilises around a stationary length. If this stationary length is large, the excursions around it can be described by a CIR process, and depletions occur as a Poisson process.

(v) In the limit of large volumes and/or small increments, the queue dynamics can be described in a continuous setting using a Fokker–Planck equation, which can be solved analytically using standard PDE tools.

5.6 Further Reading

General

Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. 2). John Wiley and Sons.

Van Kampen, N. G. (1983). *Stochastic processes in physics and chemistry*. North-Holland.

Risken, H. (1984). *The Fokker–Planck equation*. Springer, Berlin-Heidelberg.

Chen, H., & Yao, D. D. (2001). *Fundamentals of queueing networks: Performance, asymptotics, and optimisation* (Vol. 46). Springer Science & Business Media.

Gardiner, C. W. (1985). *Stochastic methods*. Springer, Berlin-Heidelberg.

Redner, S. (2001). *A guide to first-passage processes*. Cambridge University Press.

Whitt, W. (2002). *Stochastic process limits*. Springer-Verlag.

Oksendal, B. (2003). *Stochastic differential equations: An introduction with applications*. Springer.

Abergel, F., Chakraborti, A., Anane, M., Jedidi, A., & Toke, I. M. (2016). *Limit order books*. Cambridge University Press.

Queue Models for LOB Dynamics

Cont, R., Stoikov, S., & Talreja, R. (2010). A stochastic model for order book dynamics. *Operations Research*, 58(3), 549–563.
Avellaneda, M., Reed, J., & Stoikov, S. (2011). Forecasting prices from Level-I quotes in the presence of hidden liquidity. *Algorithmic Finance*, 1(1), 35–43.
Cont, R. (2011). Statistical modelling of high-frequency financial data. *IEEE Signal Processing Magazine*, 28(5), 16–25.
Cont, R., & De Larrard, A. (2012). Order book dynamics in liquid markets: Limit theorems and diffusion approximations. https://ssrn.com/abstract=1757861.
Cont, R., & De Larrard, A. (2013). Price dynamics in a Markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1), 1–25.
Garèche, A., Disdier, G., Kockelkoren, J., & Bouchaud, J. P. (2013). Fokker-Planck description for the queue dynamics of large-tick stocks. *Physical Review E*, 88(3), 032809.
Guo, X., Ruan, Z., & Zhu, L. (2015). Dynamics of order positions and related queues in a limit order book. https://ssrn.com/abstract=2607702.
Huang, W., Lehalle, C. A., & Rosenbaum, M. (2015). Simulating and analyzing order book data: The queue-reactive model. *Journal of the American Statistical Association*, 110(509), 107–122.
Huang, W., & Rosenbaum, M. (2015). Ergodicity and diffusivity of Markovian order book models: A general framework. arXiv preprint arXiv:1505.04936.
Muni Toke, I. (2015). The order book as a queueing system: Average depth and influence of the size of limit orders. *Quantitative Finance*, 15(5), 795–808.
Muni Toke, I. (2015). Stationary distribution of the volume at the best quote in a Poisson order book model. arXiv preprint arXiv:1502.03871.
Lakner, P., Reed, J., & Stoikov, S. (2016). High frequency asymptotics for the limit order book. *Market Microstructure and Liquidity*, 2(01), 1650004.

Queuing Theory

Harrison, J. M. (1978). The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, 10(04), 886–905.
Abate, J., & Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing*, 7(1), 36–43.
Abate, J., & Whitt, W. (1999). Computing Laplace transforms for numerical inversion via continued fractions. *INFORMS Journal on Computing*, 11(4), 394–405.

CIR Processes

Cox, J. C., Ingersoll Jr, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, 53, 385–407.
Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2), 327–343.
Göing-Jaeschke, A., & Yor, M. (2003). A survey and some generalisations of Bessel processes. *Bernoulli*, 9(2), 313–349.

Fokker–Planck Equation and First Passage Times

Hänggi, P., Talkner, P., & Borkovec, M. (1990). Reaction-rate theory: Fifty years after Kramers. *Reviews of Modern Physics*, 62(2), 251.