Collaborating queues: large service network and a limit order book

Elena Yudovina

Emmanuel College
University of Cambridge
Submitted April 2012, revised May 2012

This dissertation is submitted for the degree of
Doctor of Philosophy

Summary

We analyse the steady-state behaviour of two different models with collaborating queues: that is, models in which "customers" can be served by many types of "servers", and "servers" can process many types of "customers".

The first example is a large-scale service system, such as a call centre. Collaboration is the result of cross-trained staff attending to several different types of incoming calls. We first examine a load-balancing policy, which aims to keep servers in different pools equally busy. Although the policy behaves order-optimally over fixed time horizons, we show that the steady-state distribution may fail to be tight on the diffusion scale. That is, in a family of ever-larger networks whose arrival rates grow as $O(r)$ (where r is a scaling parameter growing to ∞), the sequence of steady-state deviations from equilibrium scaled down by $r^{-1/2}$ is not tight. We then propose a different policy, for which we show that the sequence of invariant distributions is tight on the $r^{1/2+\epsilon}$ scale, for any $\epsilon > 0$. For this policy we conjecture that tightness holds on the diffusion scale as well.

The second example models a *limit order book*, a pricing mechanism for a single-commodity market in which buyers (respectively sellers) are prepared to wait for the price to drop (respectively rise). We analyse the behaviour of a simplified model, in which the arrival events are independent of each other and the state of the limit order book. The system can be represented by a queueing model, with "customers" and "servers" corresponding to bids and asks; the roles of customers and servers are symmetric. We show that, with probability 1, the price interval breaks up into three regions. At small (respectively large) prices, only finitely many bid (respectively ask) orders ever get fulfilled, while in the middle region all orders eventually clear. We derive equations which define the boundaries between these regions, and solve them explicitly in the case of iid uniform arrivals to obtain numeric values of the thresholds. We derive a heuristic for the distribution of the highest bid (respectively lowest ask), and present simulation data confirming it.

Contents

CHAPTER 1

Introduction

In this thesis we discuss two examples of queueing models, with the following unifying characteristics. We have interactions of two "genders" of agents; for example, customers and servers, bids and asks, or passengers and taxis. Activity (often, service) may happen when agents of opposite gender meet. However, agents of both genders come in multiple types, and quality of possible matchings depends on the pair of types being matched; some pairings may be outright impossible. To use the passenger-taxi example, a passenger may have a certain quantity of luggage, for which a taxi may not have enough room; here some pairings are impossible. Our interest will be in algorithms that determine which of the possible pairings of agents actually occur.

There is a rich body of literature discussing problems of similar flavour. We give a broader historical introduction in this chapter, mentioning current research in the chapters to which it is applicable.

The concept of the two-sided queue, and in particular of the ubiquitous taxi-stand analogy, dates at least to Kendall. Kendall [1951] briefly considers the two-sided queue with exponential arrivals, describing its distribution as the difference of two Poisson processes or a symmetric random walk. Slightly later, Brigham [1955] discusses a many-server system, thinking of the waiting periods of the servers (who here are attendants behind the counter) as well as of the customers. The paper of Foster [1959] discusses a manufacturing queue with a finite amount of waiting room; the finite buffer size implies a certain duality between the arriving jobs and the servers working on them. This duality means that the roles of the two can be interchanged without altering the mathematical analysis.

Manufacturing systems naturally lead to the concept of multiclass queues: that is, there will be stations which can be working on customers of different types, and decisions will need to be made about which customer to serve first. Customers may even revisit a station more than once. Jackson [1957] considers a manufacturing system with M "departments" containing machines; each arriving job needs to be processed by one or more of the machines, in a fixed order (which may be different for different job types). Consequently, the machines will need to choose the queue from which they are currently taking jobs. Generalizations of this set-up have been successfully considered by Jackson [1963] (in the paper which introduced the concept of "Jackson networks"), Kelly [1976] ("Kelly networks"), and Baskett et al. [1975] ("BCMP networks"). These models have a steady-state distribution, which can be computed explicitly and depends only on the mean arrival rates and mean service times of customers. Interest in such networks was renewed after Kumar and Seidman [1990] and others ([Rybko and Stolyar, 1992, Bramson, 1994, Dumas, 1997]) constructed examples of processing networks with counterintuitive stability conditions. The focus in this line of research has been on the interactions between job types that result from the multi-stage processing inherent in a manufacturing system.

Conversely, there are many queueing models in which customers require processing only once, and then leave the network; this is the natural assumption when customers are human. Kendall [1951] includes in his discussion the notion of parallel-service systems, in which there is a single (undifferentiated) server pool, and customers simply go to the next available server. A more interesting case is when customers are required to pick

a server on arrival, causing several parallel queues. Vvedenskaya et al. [1996] presents a spectacular example of a state-dependent routing algorithm improving performance of the system. Specifically, if a customer, on arrival, may look at a randomly chosen pair of queues and always enters the shorter one, then the probability of a queue size being substantially larger than average decays much faster – superexponential decay – than if the customer were simply routed to a randomly chosen queue.

Our interest is in models where both "customers" and "servers" have a choice. Interest in such models appears to be more recent. Wein [1991] introduces a model for a network with both routing decisions (the customer, or job, may choose the server from which it receives service) and scheduling decisions (the server may choose which customer to take into service next). The model in [Wein, 1991] arises as a version of the manufacturing set-up, but one in which there are several parallel lines of machines that could work on a job, and the job may choose to switch from one line to another. Kelly and Laws [1993] discuss the emergence of *resource pooling* in some models with customer routing: although there are many server pools with differentiated skills, under certain conditions they behave as if there were just a single, large, server pool. (That is, the queue size scales as it would for a single, faster, pool of servers, not as it would for a set of parallel queues.) This means that efficiency can be increased by merging smaller systems together, creating larger, more flexible working pools. The paper of van Mieghem [1995] discusses an optimal dynamic control policy for a system with multiple customer classes waiting to be serviced by a single pool of servers.

A common theme in analysis of queueing networks is using some form of asymptotic approximations for analysing "large" networks. This is because, with few exceptions, computing the steady-state distribution of a particular network is difficult. Moreover, even when the distribution can be written down in closed form, for example in [Kelly, 1976], the answer is often unenlightening. In practice, asymptotics which reveal the scaling behaviour ("If I double the arrival rates and number of servers, what will happen to the queue size?") are often more useful. Several scaling regimes are commonly used, among them "heavy traffic" (introduced by Kingman [1962]) and "diverse routing" (studied by Ziedins and Kelly [1989]; see also Whitt [1985]). Often a single queueing model can support several limiting regimes. For example, Halfin and Whitt [1981] made the interesting observation that in a system with many servers, the customer waiting times could be kept small even when the load on the system was quite high. This is a "heavy traffic" scenario which cannot be observed in the conventional heavy traffic scaling. (Conventional heavy traffic assumes that, as arrival rates increase, the service rates of individual servers increase proportionally; here we instead have more servers working at the original speed.) This implies that in a large service system (such as a call centre), the overstaffing necessary for all customers to have small waiting times is much smaller than would be expected from conventional heavy traffic approximations. There is also a flourishing theory of diffusion approximations (see for instance [Harrison and Nguyen, 1990]), which studies the scaling behaviour of the stochastic process of deviations of the system from some nominal working point, approximating it by the solution of an appropriate stochastic differential equation. Throughout this thesis, we will frequently be interested in asymptotic questions.

Frequently there is a tension between the limiting regime imposed by the steady-state behaviour of the system (i.e., $t \to \infty$), and the limiting regime imposed by taking a "large" system. (Large here means that we consider a family of systems, indexed by $r \to \infty$; for some scaling parameter r; typically, r determines the rate at which work arrives into the r^{th} system at rate λr). In particular, in situations when one is interested in the long-term behaviour of a large system, one could consider taking these limits in either order;

and much interesting research has been concerned with the question of whether the two procedures commute. In terms of the diagram below, we would like to know whether there is convergence along all edges, and if so, whether the limits can be taken in either order.

$$\begin{array}{ccc} \hat{X}^r(t) & \xrightarrow[t\to\infty]{\text{steady-state distribution}} & \hat{X}^r \\ \text{limiting process}\Big\downarrow r\to\infty & & \text{appropriate scaling?}\Big\downarrow r\to\infty \\ \hat{X}(t) & \xrightarrow[t\to\infty]{\text{steady-state distribution?}} & \hat{X} \end{array}$$

Much of Chapter 2 is concerned with this question explicitly. In Chapter 3, although we are unable to prove the existence of a steady-state distribution, we certainly will be interested in the question of whether different asymptotic approximations commute.

Last, we note that the problem of interactions between two "genders" of agents, which we informally posed, does not have to be modelled as a queueing system. That is, we may not want to introduce a stochastic process of arrivals and of service times. For example, Caldentey et al. [2009] study the problem using so-called *infinite bipartite matchings*. Specifically, they make the assumption that, in the problem of pairing "passengers" and "taxis", there is an infinite stream supplying each type of agent, and the goal of each agent is simply to find a match. The Caldentey et al. [2009] model was inspired by housing projects, in which interested applicants are matched with housing as it becomes available, and there may not be a meaningful notion of an arrival process, not to mention service time. Our model of the limit order book in Chapter 3 fits into their framework.

Acknowledgments

This thesis would have been impossible without two people: first, my supervisor at the University of Cambridge, Professor Frank P. Kelly, and second, Dr. Alexander L. Stolyar at Alcatel-Lucent Bell Labs. My deepest gratitude is due to both of them.

Some of the many other people to conversations with whom portions of this thesis can be traced are Maury Bramson, Michael Chmutov, Sergey Foss, Florian Simatos, Yuri Suhov, Neil Walton, Richard Weber, Damon Wischik, and the referee of [Stolyar and Yudovina, 2010] for the *Annals of Applied Probability*. Special thanks to Daniel Whalen for help with the Mathematica graphics. Many thanks also to Michael Chmutov, Vladimir Dokchitser, and Daniel Whalen for proofreading parts of the dissertation.

My PhD studies have been generously funded by the US National Science Foundation Graduate Research Fellowship.

Last but not least, I am indebted to the generosity of colleagues, friends, and family members, who have surrounded me with patience, kindness, and biscuits throughout the thesis-writing process.

Map of thesis

Each chapter is self-contained, with its own introduction and summary of recent relevant literature.

In Chapter 2 we introduce the model of a large service network, motivated by call centres. We study two algorithms for making routing and scheduling decisions. One (LQFS-LB) arises naturally from a static planning problem, but we show that it can lead to undesirable behaviour (unstable fluid-scale approximations over finite time horizon, and "large" steady-state deviations from equilibrium). The other (LAP) is designed to squash such instability, and we prove that the steady-state deviations from equilibrium

when LAP is used are "not too large". We discuss finite-time horizon behaviour on a variety of scales, steady-state behaviour, and the interplay between them.

In Chapter 3 we discuss the concept of limit order books, and formulate a simple, analytically tractable model of it. We then show that even such a simple model can have interesting behaviour.

Appendices contain extra information. Appendix A gives a brief overview of the various notions of continuity of functions that we use in the thesis. Appendix B contains a discussion of unstable networks, and offers an intuition for wanting to consider trees in Chapter 2. Appendix C presents a summary of the results in Halfin and Whitt [1981], which form the inspiration and basis for much of Chapter 2. Appendix D contains computations which would be too bulky to include in the main text.

Assumptions, definitions, examples, lemmas, propositions, and theorems share a common numbering, and are numbered consecutively within each chapter; equations are numbered sequentially throughout the thesis. Except in this chapter and the appendices, section numbers refer to sections of the same chapter.

Pages 89, 93, 94, and 117 are best viewed in colour.

Location of original results

The following sections contain new models, algorithms, or results: 2.3–8, 3.1–11, Appendix D.

Chapter 2 presents work undertaken in collaboration with Alexander L. Stolyar (Alcatel-Lucent Bell Labs, Murray Hill, NJ). The results follow [Stolyar and Yudovina, 2010], [Stolyar and Yudovina, 2012] and the corrections suggested by the referees of [Stolyar and Yudovina, 2010] for the *Annals of Applied Probability*, but the proofs in many sections (particularly §2.7) have been expanded. The associated computations in Appendix D.1–3 are my own, and the expository Appendices B and C are new in the thesis.

Although there necessarily is a certain amount of overlap between the material in this dissertation and the fourth term report and the Smith-Knight and Rayleigh-Knight prize essay I submitted at the end of my fourth term at Cambridge, essentially all of the text has been rewritten.

Notation

Vectors and matrices. In Chapter 2, we will encounter many vectors indexed by sets $\mathcal{I}$, $\mathcal{J}$, $\mathcal{E}$, $\mathcal{C}(j)$, and $\mathcal{S}(i)$. $\mathcal{I}$ and $\mathcal{C}(j)$ index customer types; their elements are denoted i, i', etc. $\mathcal{J}$ and $\mathcal{S}(i)$ index server types; their elements are denoted j, j', etc. $\mathcal{E}$ indexes (a subset of) customer-server type pairings; its elements are denoted (ij).
For any symbol γ,

$$(\gamma_i, i \in \mathcal{I}) = (\gamma_i) = \gamma_{\mathcal{I}}.$$

Similarly, $(\gamma_j, j \in \mathcal{J}) = (\gamma_j) = \gamma_{\mathcal{J}}$, and $(\gamma_{ij},\ (ij) \in \mathcal{E}) = (\gamma_{ij}) = \gamma_{\mathcal{E}}$. Occasionally, we also use $\gamma_{\mathcal{I}\mathcal{J}} = (\gamma_{ij}, i \in \mathcal{I}, j \in \mathcal{J})$.
Although elements of $\gamma_{\mathcal{E}}$ may have a double index ij, we treat $\gamma_{\mathcal{E}}$ as a (column) vector, not as a matrix.
Unless specified otherwise,

$$\sum_j \gamma_{ij} = \sum_{j \in \mathcal{S}(i)} \gamma_{ij}, \qquad \sum_i \gamma_{ij} = \sum_{i \in \mathcal{C}(j)} \gamma_{ij}.$$

The symbols γ and Γ will reappear as placeholders, but do not have any specific meaning in the thesis.

In matrix expressions, vectors are column vectors unless specified otherwise. For a (column) vector v, its transpose (a row vector) is denoted $v^\top$. Similarly, for a matrix M, its transpose is denoted $M^\top$.
For a vector $v \in \mathbb{R}^d$, its Euclidean norm is denoted $\|v\|$.
The zero vector is denoted simply 0; it will be clear from context that the quantity is a vector.

Sets. $\mathcal{M}$, $\mathcal{N}$ are manifolds.
$\mathcal{I}$, $\mathcal{J}$, $\mathcal{E}$, $\mathcal{C}(j)$, $\mathcal{S}(i)$ are discrete index sets.
$\mathcal{A}$ is an event.
(Also written in the same script, but not sets: $\mathcal{P}$ is a partial ordering; $\mathcal{L}$ is a Lyapunov function; $\mathcal{F}$ and $\mathcal{F}_n$ are σ-algebras.)
The one-point compactification of $\mathbb{R}^d$ is denoted $\overline{\mathbb{R}^d}$. The σ-algebra on $\mathbb{R}^d$ and $\overline{\mathbb{R}^d}$ is always the Borel σ-algebra.
The space of RCLL functions with domain $[\eta, \infty)$ and values in $\mathbb{R}^d$ is denoted $D^d[\eta, \infty)$. Usually, $\eta = 0$. (RCLL means "right-continuous with left limits", see below under Functions.) The notion of convergence on $D^d[\eta, \infty)$ is uniform convergence on compact sets; see below under Convergence.

Measures. Measures on $\mathbb{R}^d$ for the appropriate dimension d, or on its one-point compactification $\overline{\mathbb{R}^d}$, are denoted using Gothic script; e.g., $\mathfrak{M}$, $\mathfrak{A}$, $\mathfrak{D}$, $\mathfrak{Q}$. ($\mathfrak{A}$ is meant to resemble "A"; $\mathfrak{Q}$ is meant to resemble Q.) Of these, $\mathfrak{A}$, $\mathfrak{D}$, and $\mathfrak{Q}$ are counting measures on $\mathbb{R}$.
For a measure $\mathfrak{M}$ on $\mathbb{R}$, we write $\mathfrak{M}[a, b]$, $\mathfrak{M}[a, b)$, and $\mathfrak{M}\{a\}$ to denote $\mathfrak{M}([a, b])$, $\mathfrak{M}([a, b))$, and $\mathfrak{M}(\{a\})$ respectively.
π and ϖ are probability measures on $[0, 1]$.

Partial orderings. Partial orderings are named $\mathcal{P}$ and variations thereon, and denoted $x \prec y$.
If x and y are incomparable, i.e. none of $x \prec y$, $x = y$, or $x \succ y$ holds, we write $x \sim y$.
(Our use of $\preceq$ in §2.5 has nothing to do with partial orderings.)

Functions and random processes. For functions (or random processes) $(\gamma(t),\ t \geq 0)$ we often write $\gamma(\cdot)$; we also do this for functions with domain different from $[0, \infty)$. For a vector of functions, we may combine the shorthand vector notation with the shorthand function notation: for example, $(\gamma_i(\cdot))$ and $\gamma_{\mathcal{I}}(\cdot)$ both signify $((\gamma_i(t), i \in \mathcal{I}),\ t \geq 0)$.
For γ_t a state variable indexed by time, $\gamma_{t^-} \equiv \lim_{\epsilon \downarrow 0} \gamma_{t-\epsilon}$ and $\gamma_\infty \equiv \lim_{t \uparrow \infty} \gamma_t$, provided the limit exists.
The indicator function of a set A is denoted $\mathbf{1}_A$; that is, $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$, and 0 otherwise.
The symbol $\mathcal{L}$ denotes a Lyapunov function; see §2.6-7 and §2.8.
The term "RCLL" means "right-continuous with left-limits" (also denoted càdlàg in literature). These are functions $\gamma(\cdot)$ for which $\gamma(t^-)$ exists but need not be equal to $\gamma(t)$, and $\gamma(t^+) \equiv \lim_{\epsilon \downarrow 0} \gamma(t + \epsilon)$ exists and is equal to $\gamma(t)$.
The derivative of a function $f(\cdot)$ is denoted $\dot{f}$.

Convergence. The symbol $\Longrightarrow$ denotes convergence in distribution of random processes in the Skorohod space $D^d[\eta, \infty)$, uniformly on compact sets[1].
The symbol $\xrightarrow{w}$ denotes weak convergence of probability measures on $\mathbb{R}^d$ or its one-point

[1]The usual topology on the space $D^d[\eta, \infty)$ of RCLL functions is the "Skorohod topology," or more precisely one of the Skorohod topologies. The need for a topology other than one of uniform convergence arises because $D^d[\eta, \infty)$ is not separable in the topology of uniform convergence on compact sets. However, differences between convergence in the uniform sense and convergence in the Skorohod sense arise only at jump points of the limiting process, and all limiting processes we consider will be continuous. There is an excellent discussion of this point – and the Skorohod topology – in [Pollard, 1984, Chapter VI.1].

compactification $\overline{\mathbb{R}^d}$. It also denotes convergence in law of the associated random variables.
The symbol $\to$ denotes ordinary convergence in $\mathbb{R}^d$, $\overline{\mathbb{R}^d}$, or $D^d[\eta, \infty)$.
The term *u.o.c.* means *uniform(ly) on compact sets*; the domain may be defined explicitly, or be obvious from the context.
The term *w.p.1* means *with probability 1*, which is the same as *almost surely*.

Miscellaneous. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the greatest integer less than or equal to x. (The notation $[x]$ in §3.6 is unrelated.)
We use $\equiv$ as the assignment operator, and $=$ in equalities. That is, if we define $x \equiv 2+2$, then the equality $x = 4$ holds.
The index r is a scaling parameter. We are typically interested in the behaviour of quantities as $r \to \infty$. For a function $f(r)$ we say that $f(r)$ is $O(r)$ if $|r^{-1}f(r)|$ is bounded as $r \to \infty$, and $f(r)$ is $o(r)$ if $|r^{-1}f(r)| \to 0$ as $r \to \infty$. We write $o(1)$ to mean some function which converges to 0. For a sequence of random variables, *iid* means *independent and identically distributed*.

CHAPTER 2

Large service network

Introduction

In this chapter we model a system in which service requests of several different types arrive externally, are processed by servers with varying skills, and leave the system. Examples of such systems include call centres, cloud computing (where jobs submitted to the cloud take the role of service requests, and the machines take the role of servers; instead of "skill" the servers may be differentiated by available memory and processing power), as well as emergency wards in hospitals. Our primary example will be a call centre; we will therefore refer to the service requests as "customers".

A common feature of these applications is the large number of servers they employ, and the relatively unscalable processing requirements for each service activity. (A call centre agent may get marginally faster at processing calls on a busy day, but the effect is unlikely to be significant.) To compensate for the inflexible speed of servers, we instead may adjust their number (e.g., by hiring more call centre agents). To gain insight into the behaviour of a large call centre, we will be looking at the "many-server" asymptotic regime, in which the individual contribution of each server to the total processing capacity becomes negligible as the system grows. When the arrival rate of calls is close to the maximal that the servers are capable of processing, this asymptotic approximation is also known as the "Halfin-Whitt regime", after the authors of [Halfin and Whitt, 1981]. Halfin and Whitt [1981] show that, for a model of a single-class many-server queue, by carefully managing staffing levels as the system grows larger, the expected time that customers spend queueing prior to entering service while the probability that an arriving customer has to wait will tend to a constant strictly between 0 and 1. This is achieved by having a system with $O(r)$ arrival rate, putting in $O(\sqrt{r})$ extra servers beyond the minimal number necessary to process all the arriving work on average. (In contrast, usual heavy load techniques only guarantee small customer waiting times – and then asymptotic probability 0 of having to wait at all – with $O(r)$ overstaffing. The difference is considerable when there are hundreds – or, in the case of a large call centre, thousands – of agents.)

The analysis of Halfin and Whitt (which we briefly summarize in Appendix C) uses undifferentiated customers and servers. However, in a call centre there are many types of customer requests (e.g., "I lost my credit card," "I can't log into online banking," and "I need to transfer money to an account overseas"[1]), which are typically serviced by different pools of agents. The different pools are not entirely separated, because agents are typically cross-trained: for example, although we have assumed that "online banking" and "lost credit card" are two different call types, there probably are agents who can both email you a password reminder and block your lost credit card. It is likely that not every server can service every request type, and the associated service times may well vary. The challenge for a call centre then becomes to assign customers to servers in such a way that the entire system "looks like" a single pool of agents; in particular, so that the entire

[1]The examples of call types listed here are simply guesses formulated while waiting on the phone; the actual division of incoming requests into classes in a bank's call centre could be completely different.

system has little customer waiting with only $O(\sqrt{r})$ overstaffing, even if the arrival rates of calls of different types change.

Let us classify the servers by the training they have received, and the associated average speed with which they can process different types of customer requests. If all of the parameters of the system, such as arrival rates and mean service times, are known, we can use the solution of the *static planning problem* (§1.2) to design a simple probabilistic routing mechanism. If the parameters of the system satisfy the *complete resource pooling* condition (Assumption 2.4), Halfin-Whitt-like behaviour is likely to emerge. (We discuss this point in §1.2, after Assumption 2.4.) However, in practice we usually want algorithms which do not rely on precise knowledge of arrival rates, since these are external to the system and may well change. In this case, somewhat less is known. A myopic, "maximal weight"-like policy, which is optimal when the number of servers is fixed [Mandelbaum and Stolyar, 2004], is known to not have optimal overstaffing requirements in the many-server regime [Stolyar and Tezcan, 2010]. Stolyar and Tezcan [2010, 2011] propose a "shadow routing" algorithm, which they conjecture *does* have optimal overstaffing behaviour.

Both the maximal weight and the shadow routing algorithms rely on the precise knowledge of the mean service times. However, in real systems these also can only be approximated. It would be preferable if the algorithm for assigning jobs to servers used only the information on system state (such as queue sizes and number or proportion of idle servers in each pool) and did *not* explicitly rely on either the arrival or the service rates. The two algorithms we investigate in this chapter rely only on knowing the *basic activity tree*. This is defined in §1.2; intuitively, it indicates the set of "most efficient" customer-server type pairings for the given arrival pattern and set of service rates. We would expect it to change only rarely, because computing the basic activity tree only requires approximate knowledge of the system parameters, and is insensitive to small perturbations.

We will consider two algorithms. One (longest-queue freest-server load balancing, or LQFS-LB) has a more natural definition; however, we show that it can "misbehave", in the sense of having large deviations from equilibrium (which will be defined later). In particular, we show that almost always in steady state the system is too far from equilibrium for diffusion approximations to be applicable. (The finite-time-horizon diffusion approximation for a family of algorithms including this one has been rigorously constructed by Gurvich and Whitt [2009]; we summarize the relevant results in §7.1.) We informally conjecture that this behaviour is "rare": all the counterexamples we have been able to construct have somewhat unrealistic parameter values. We show that for certain parameter values the algorithm really does show the Halfin-Whitt regime behaviour (infinitesimal average waiting times and finite probability of customer waiting, with $O(\sqrt{r})$ overstaffing).

The other algorithm we consider (leaf activity priority, or LAP) is more robust, but its operating point is less intuitive. For it, we conjecture the correct overstaffing behaviour (infinitesimal average waiting times and finite probability of customer waiting, with $O(\sqrt{r})$ overstaffing). We prove a slightly weaker result, namely that the deviations of the system state from equilibrium are $O(r^{1/2+\epsilon})$ for any $\epsilon > 0$.

As was mentioned in Chapter 1, for queueing models in general, and many-server models in particular, there is a tension between the time scaling and space scaling. We will be interested in the long-term behaviour of a large network, and we will consider a family of ever-larger networks, indexed by a scaling parameter r. We might then do one of two things: (a) consider the associated family of steady-state distributions (possibly, centered and rescaled), and take the limit; or (b) construct a limiting process which approximates system behaviour (appropriately scaled) over a finite time horizon, and

take its steady-state distribution. Schematically,

$$\begin{array}{ccc} \hat{X}^r(t) & \xrightarrow[t\to\infty]{\text{steady-state distribution}} & \hat{X}^r \\ \text{limiting diffusion process}\Big\downarrow r\to\infty & & ?\Big\downarrow r\to\infty \\ \hat{X}(t) & \xrightarrow[t\to\infty]{\text{steady-state distribution?}} & \hat{X} \end{array}$$

Typically, in this diagram it is "easier" to go down and across, i.e. to the limiting process and then to its steady-state behaviour. There are standard techniques for proving convergence of the, appropriately scaled, state of a large queueing network to a Semi-Martingale Reflected Brownian Motion (SRBM); this was done in conventional heavy traffic by Harrison and Williams [1987], and for multi-server models examples can be found in [Mandelbaum et al., 1998, Pang et al., 2007]. There is also a large body of work studying when the approximating SRBM has an invariant distribution. While the full characterisation of the necessary and sufficient conditions for the SRBM to have an invariant distribution has not been accomplished[2], a large class of sufficient conditions is known. (The papers [Harrison and Williams, 1987, El Kharroubi et al., 2000, 2002, Bramson et al., 2010] collectively characterise stability in at most three dimensions. Perhaps more relevantly for queueing applications, [Harrison and Williams, 1987] provides a set of sufficient conditions, which works in arbitrary dimension and is more natural in the context of queueing models.)

However, the diagram above need not commute, primarily because the family of invariant distributions $\{X^r\}$ need not be tight (so need not have any limit points as $r \to \infty$). In the many-server setting, this *limit interchange problem* has been particularly challenging. While there are a few individual results (notably, Corollary 2 in Halfin and Whitt [1981], reproduced in Appendix C as Theorem C.4; more recently, Gamarnik and Zeevi [2006], Gamarnik and Momcilovic [2008], Gamarnik and Stolyar [2012] as well as Stolyar and Yudovina [2010]), they show tightness only in very specialised settings. In the framework of multitype one-hop queueing networks, we provide an example of a situation where diffusion-scale tightness holds (in §7.4), as well as an example of a natural algorithm for which it does not hold (in §7.2). For the case of the leaf activity priority, in §8 we prove a family of tightness results on scales bigger than the diffusion scale (which are rarely encountered in the literature), and state a conjecture a tightness result on the diffusion scale.

There is a third direction from which the algorithms we consider look interesting. An important aspect of queueing theory is the study of stability of queueing models. For certain types queueing networks (Jackson networks [Jackson, 1963], Kelly networks [Kelly, 1976], BCMP networks [Baskett et al., 1975]), the stability criterion is simple: if none of the servers are on average receiving more jobs than they can process, then the network is stable. The natural conjecture that this is the only requirement for network stability was essentially disproved[3] in 1990 with the Kumar-Seidman network [Kumar and Seidman, 1990]; several other examples (e.g., [Rybko and Stolyar, 1992], [Bramson, 1994], [Dumas, 1997]) have since been produced. A common feature of most of these examples of instability is that there is a certain loop, or cycle, in the structure of the job flow graph, and after "going around" this loop the number of unserviced jobs in the system increases.

[2]In fact, as Gamarnik and Katz [2010] show, a simple set of necessary and sufficient conditions *cannot* be identified: the question of whether a SRBM is positive recurrent is undecidable.

[3]There wasn't a formal conjecture to disprove; rather, the series of papers [Kumar and Seidman, 1990], [Rybko and Stolyar, 1992], [Bramson, 1994], [Dumas, 1997] and others exhibited increasingly natural disciplines displaying instability without any given station being overloaded.

The curious feature of the algorithm we analyse is that, although its routing graph is constrained to be a *tree* (i.e., cycle-free), it nevertheless supports unstable, exponentially growing perturbations. We discuss this further in Appendix B.

1. Call centre model and the static planning problem

1.1. Scaling regime and state descriptor. Consider a queueing model in which there are I customer classes, or types, labelled $1, 2, \ldots, I$, and J server (agent) pools, or classes, labelled $1, 2, \ldots, J$. Generally, we will use the subscripts i, i' (and sometimes k) for customer classes, and j, j' (and sometimes k) for server pools; the sets of customer classes and server classes will be denoted by $\mathcal{I}$ and $\mathcal{J}$ respectively.

We are interested in the scaling properties of the system as it grows large. The meaning of "grows large" is as follows. We consider a sequence of systems indexed by a scaling parameter $r \to \infty$. As r grows, the arrival rates and the sizes of the service pools, but not the speed of service, increase. Specifically, in the rth system, customers of type i enter the system as a Poisson process of rate $\lambda_i^r = r\lambda_i + o(r)$, while the jth server pool has $r\beta_j$ individual servers. (All λ_i and β_j are positive parameters.)

We model the system as *input-queued*. That is, customers are only assigned to a server type when they are taken into service; if queueing occurs, there is a separate queue for each customer type. We do not allow customers to abandon the system before being served. (In this chapter we will be discussing a system in underload or in Halfin-Whitt-type heavy traffic; for it, waiting times ought to be negligible, and abandonment should not be important.) When a customer of type i is accepted for service by a server in pool j, the service time is exponential of rate μ_{ij}; the service rate depends both on the customer type and the server type, but *not* on the scaling parameter r. If customers of type i cannot be served by servers of class j, the service rate is $\mu_{ij} = 0$. All interarrival and service times are taken to be independent exponentials.

We present a schematic diagram of such a model in Figure 2.1.

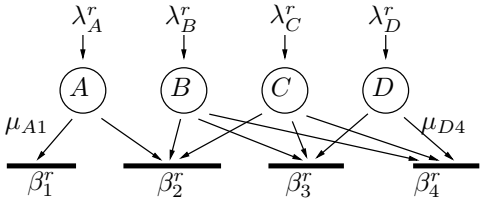

FIGURE 2.1. Schematic diagram of a queueing system showing the arrival rates, service rates, and number of servers in each pool. The absence of an edge implies the corresponding service rate is zero; e.g., here $\mu_{B1} = 0$.

For the system with scaling parameter r, we introduce the following notation for the system state at time t:
$\Psi_{ij}^r(t)$ is the number of servers of type j serving customers of type i;
$\Psi_j^r(t) \equiv \sum_{i\in\mathcal{I}} \Psi_{ij}^r(t)$ is the total number of busy servers of type j;
$\Psi_i^r(t) \equiv \sum_{j\in\mathcal{J}} \Psi_{ij}^r(t)$ is the total number of servers serving type i customers;
$P_j^r(t) \equiv \Psi_j^r(t)/\beta_j$ is the instantaneous load of server pool j;
$Q_i^r(t)$ is the number of customers of type i waiting for service;
$X_i^r(t) \equiv \Psi_i^r(t) + Q_i^r(t)$ is the total number of customers of type i in the system;
$-Z_j^r(t) \equiv \beta_j r - \Psi_j^r(t)$ is the number of idle servers of type j (note that $Z_j^r(t) \le 0$).

We further describe the state of the routing choices that have been made by the algorithms up to time t:
$A_i^r(t)$ is the total number of customers of type i that have arrived into the system in the

interval $[0,t]$;
$D_{ij}^r(t)$ is the total number of customers of type i that completed service in pool j (and departed the system) in the interval $[0,t]$;
$\Xi_{ij}^r(t)$ is the total number of customers of type i that entered service in pool j in the interval $[0,t]$.

1.2. Static Planning Problem. The load-balancing objective is to minimize the maximal proportion of occupied servers of any given type. Suppose that the rth system has a well-defined average rate λ_{ij}^r at which requests of type i are sent to servers of type j. Intuitively, a load-balancing algorithm should aim to have $\lambda_{ij}^r \approx \lambda_{ij} r$, where $\{\lambda_{ij}\}$ is an optimal solution to the following *static planning problem* (SPP) (see [Harrison, 2000]):

$$\min_{\lambda_{\mathcal{IJ}},\rho} \rho, \tag{1a}$$

subject to

$$\lambda_{ij} \geq 0, \ \forall i,j \tag{1b}$$

$$\sum_{j\in\mathcal{J}} \lambda_{ij} = \lambda_i, \ \forall i \tag{1c}$$

$$\sum_{i\in\mathcal{I}} \lambda_{ij}/(\beta_j\mu_{ij}) \leq \rho, \ \forall j. \tag{1d}$$

Definition 2.1. The optimal value of ρ in (1) is called the *load* on the system. If $\rho < 1$, the system is called *underloaded*; if $\rho = 1$, the system is called *critically loaded*.

We will not consider the overloaded case $\rho > 1$, in which case some of the customers must abandon the system for it to be stable. Talreja and Whitt [2008] discuss fluid model asymptotics for overloaded many-server systems.

The dual problem to (1) is

$$\max_{\nu_{\mathcal{I}},\alpha_{\mathcal{J}}} \sum_i \lambda_i\nu_i, \tag{2a}$$

subject to

$$\alpha_j \geq 0, \ \forall j \tag{2b}$$

$$\sum_{j\in\mathcal{J}} \alpha_j = 1 \tag{2c}$$

$$\alpha_j \geq \nu_i\beta_j\mu_{ij}, \ \lambda_{ij}\left(\alpha_j - \nu_i\beta_j\mu_{ij}\right) = 0, \ \forall i,j. \tag{2d}$$

Definition 2.2. The optimal value of ν_i in (2) is called the *workload* associated with a job of type i. The optimal value of α_j is called the *rate at which server pool j can process workload*.

In a system indexed by r, the rate at which server pool j processes workload scales as r, whereas the workload associated with an individual job of type i does not scale.

Strong duality guarantees that

$$\sum_{j\in\mathcal{J}} \alpha_j = 1, \qquad \sum_{i\in\mathcal{I}} \lambda_i\nu_i = \rho \sum_{j\in\mathcal{J}} \alpha_j = \rho. \tag{3}$$

REMARK 2.3. The workloads ν_i and rates α_j are *not* intrinsic to the service system: they depend on the parameters $\beta_{\mathcal{J}}$, $\mu_{\mathcal{I}\mathcal{J}}$, but also on the arrival rates $\lambda_{\mathcal{I}}$. In other words, the same call centre faced with two different patterns of calls $\lambda_{\mathcal{I}}$, $\tilde{\lambda}_{\mathcal{I}}$ may well assign different values of "workload" to jobs of a given type, and different rates of processing said workload by the server pools.

However, the feasible set of the dual problem (2) defining $\nu_{\mathcal{I}}$ and $\alpha_{\mathcal{J}}$ depends only on the parameters $\beta_{\mathcal{J}}$, $\mu_{\mathcal{I}\mathcal{J}}$, which are intrinsic to the service system. Since (2) is a linear program whose optimum is always attained at one or more vertices of the feasible set, there is a finite set of possible "workloads", and this set depends only on the parameters intrinsic to the system. Moreover, provided the arrival rates $\lambda_{\mathcal{I}}$ are such that the maximum is attained at only one vertex of the dual feasible set, there will be a unique possible set of workloads, and the same set will work for all sufficiently close values $\tilde{\lambda}_{\mathcal{I}}$.

2. Complete resource pooling

Throughout this chapter, we make the following *complete resource pooling* (CRP) assumption:

ASSUMPTION 2.4. *The SPP* (1) *has a unique optimal solution*

$$\{\lambda_{ij},\ i \in \mathcal{I},\ j \in \mathcal{J}\},\ \rho.$$

The solution is such that the set of pairs, or edges, (ij) *for which* $\lambda_{ij} > 0$ *forms a (connected) tree* [4] *in the graph with vertex set* $\mathcal{I} \cup \mathcal{J}$.

The CRP assumption can equivalently be formulated as "The linear program (1) has a unique, non-degenerate solution" (and hence so does its dual). The term "complete resource pooling" was introduced in the paper of Harrison and López [1999], where this condition is used to simplify diffusion-scale analysis; but variants of the condition are ubiquitous in discussions of systems with multiple server types that may share customers.

DEFINITION 2.5. A *basic activity* is a pair (ij) such that $\lambda_{ij} > 0$ in the optimal solution to (1). The *basic activity tree* $\mathcal{E}$ is the graph formed by the (undirected) edges (ij) which are basic activities.

Assumption 2.4 consists of two parts. The assumption that optimal solution is unique and the graph formed by basic activities contains no cycles holds "generically": in systems where it is violated, the parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, and $\mu_{\mathcal{I}\mathcal{J}}$ are linked by a set of polynomial equations [Stolyar and Tezcan, 2011, Theorem 2.2]. Since the arrival rates at a call centre typically oscillate throughout the day, it seems reasonable to assume that most of the time the parameters will not be so well-matched.

The assumption that the graph is connected may well fail for a large range of parameters. If it does fail, then in heavy load it is optimal to run the system as several noninteracting subsystems, where sharing occurs within each subsystem but not between them. In this case, all of the analysis below applies to each of the connected components separately.

When Assumption 2.4 holds, $\mathcal{E}$ is also the graph formed by edges (ij) along which equalities hold in (2d). (Without the CRP assumption, there may be additional edges along which equality holds.)

In Figure 2.2 we show the optimal tree associated with a particular set of parameter values. The workloads are $\nu_A = \frac{1}{12}, \nu_B = \frac{1}{18}, \nu_C = \frac{1}{54}, \nu_D = \frac{1}{54}$, and the corresponding service rates are $\alpha_1 = \frac{1}{3}, \alpha_2 = \frac{1}{3}, \alpha_3 = \frac{1}{9}, \alpha_4 = \frac{2}{9}$.

[4] A *tree* is a connected graph without cycles. Its *leaves* are nodes with only one outgoing edge.

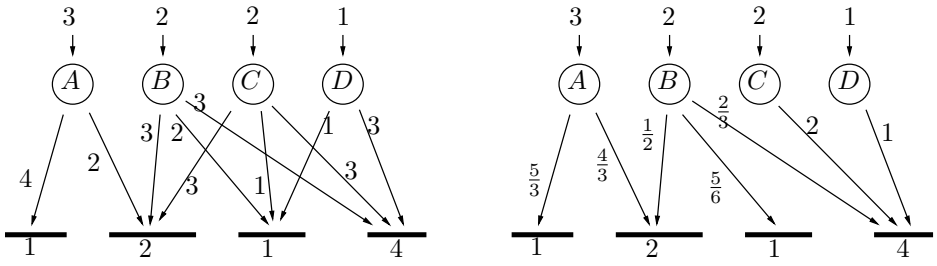

FIGURE 2.2. Sample parameters for the queueing system, and associated solution to the static planning problem ($\rho = 5/12$).

DEFINITION 2.6. For a customer type i, let $\mathcal{S}(i) \equiv \{j : (ij) \in \mathcal{E}\}$ denote the set of server types to whom customers of type i are routed in the solution to the static planning problem; for a server type j, let $\mathcal{C}(j) \equiv \{i : (ij) \in \mathcal{E}\}$ denote the set of customer classes that servers of type j process in the solution to the SPP.

We can think about workload as follows. Jobs of type i arrive at rate λ_i bringing a certain amount ν_i of work with them. A server in pool j that is working on a job of *any* type $i \in \mathcal{C}(j)$ is processing system workload at rate $\frac{1}{\beta_j^r}\alpha_j$. If it is working on a job of some type $i' \notin \mathcal{C}(j)$, then its rate of processing workload is strictly slower than $\frac{1}{\beta_j^r}\alpha_j$. Consequently, if we want to run the system efficiently, we should only assign servers to work on customers of types $i \in \mathcal{C}(j)$. On the other hand, it seems intuitively plausible that any "reasonable" policy which assigns customers to servers without straying outside the basic activity tree will result in the same behaviour of the system workload; that is, effectively we will have "merged" the server pools into a single large pool that is processing system workload as efficiently as it can.

Throughout the discussion of call centre models, we make the following additional assumption:

ASSUMPTION 2.7. *The basic activity tree $\mathcal{E}$ is known in advance. All assignments of customers to servers are made along edges of the basic activity tree.*

This will ensure that routing choices are such that $\lambda_{ij}^r = 0$ for $(ij) \notin \mathcal{E}$; that is, all servers, when they are busy, will be processing workload as quickly as they can.

REMARK 2.8. As in Remark 2.3 on the dual workload, the basic activity tree $\mathcal{E}$ depends on the arrival rates $\lambda_{\mathcal{I}}$, as well as on $\beta_{\mathcal{J}}$ and $\mu_{\mathcal{IJ}}$. If we are at a point where CRP holds (implying that there is a unique possible choice of $\mathcal{E}$), then a small perturbation in parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, or $\mu_{\mathcal{IJ}}$ will not change the tree (although it will change the optimal rates $\lambda_{\mathcal{E}}$).

This suggests that, as long as the arrival rates $\lambda_{\mathcal{I}}$ and service rates $\mu_{\mathcal{IJ}}$ are not subject to wild fluctuations, we can *separate time scales*. First, over a longer time scale we estimate the parameters enough to determine which of the possible basic activity trees is present in our case, and then over a very short time scale we route individual customers to servers based on the knowledge of the specific tree. Our discussion in what follows is concerned only with the routing of customers on the short time scales; (approximate) identification of the basic activity tree could be done either by measuring $\lambda_{\mathcal{I}}$ and $\mu_{\mathcal{IJ}}$ and solving the static planning program, or (assuming the $\lambda_{\mathcal{I}}$ are more variable than the $\mu_{\mathcal{IJ}}$) by the shadow routing algorithm of Stolyar and Tezcan [2011]. (The shadow routing algorithm will give incorrect rates $\tilde{\lambda}_{\mathcal{E}}$, but will identify the correct set $\mathcal{E} \equiv \{(ij) : \tilde{\lambda}_{ij} > 0\}$.)

3. LQFS-LB and LAP algorithms

In this section we define the two algorithms we will be considering for matching customers to servers: Longest-Queue Freest-Server Load Balancing (LQFS-LB) and Leaf Activity Priority (LAP). LQFS-LB belongs to the family of algorithms considered by Gurvich and Whitt [2009] and others (Armony and Ward [2011], Atar et al. [2011]). It is a natural routing and scheduling rule that strives to equalize the load, or proportion of busy servers, on all the server pools. LAP instead assigns static priorities to the basic activities in the basic activity tree, and strives to keep the high-priority activities "filled".

Each of the algorithms consists of two parts: routing and scheduling. "Routing" determines where an arriving customer goes if it sees available servers of several different types. "Scheduling" determines which waiting customer a server picks if it sees customers of several different types waiting in queue.

Throughout this chapter, we alternate between analysing the two algorithms. Thus, §3-4 discuss both algorithms, §5 and §7 are devoted exclusively to LQFS-LB, and §6 and §8 are devoted exclusively to LAP.

3.1. Longest-queue, freest-server load balancing algorithm (LQFS-LB). **Scheduling:** If a server of type j, upon completing a service, sees a customer of a class in $\mathcal{C}(j)$ in queue, it will pick the customer from the longest queue, i.e. $i \in \arg\max_{j\in\mathcal{C}(j)} Q_i^r$. (Ties are broken in an arbitrary Markovian manner.)

Routing: If an arriving customer of type i sees any unoccupied servers in server classes in $\mathcal{S}(i)$, it will pick a server in the least loaded server pool, i.e. $j \in \arg\min_{j\in\mathcal{S}(i)} P_j^r(t)$. (Ties are broken in an arbitrary Markovian manner.)

This algorithm is a special case of one considered by Gurvich and Whitt [Gurvich and Whitt, 2009, Remark 2.3], with constant probabilities $p_i = \frac{1}{I}$ (queues "should" be equal), $v_j = \frac{\beta_j}{\sum \beta_j}$ (the proportion of idle servers "should" be the same in all server pools). The results of [Gurvich and Whitt, 2009] which we use are briefly summarized in §7.1.

3.2. Leaf Activity Priority algorithm. The definition of Leaf Activity Priority (LAP) policy proceeds in three steps. First, we assign priorities to customer classes as follows:

(1) Pick a leaf[5] of the tree;
(2) If it is a customer class (rather than a server class), assign to it the highest priority that hasn't yet been assigned;
(3) Remove the leaf from the tree.

Without loss of generality, we assume the customer classes are numbered in order of priority (with 1 being highest).

We now assign priorities to the edges of the basic activity tree by iterating the following procedure:

(1) Pick the highest-priority customer class;
(2) Pick an edge of the activity tree going out of the class to a leaf;
(3) Assign this edge the highest priority that hasn't yet been assigned, and remove the edge;
(4) When the customer class becomes a leaf of the activity tree, assign the remaining edge out of it the highest priority that hasn't yet been assigned, and remove the edge together with the customer class.

It is not hard to verify that this algorithm will successfully assign priorities to all edges. It suffices to check that at any time the highest remaining priority customer class will have

[5] See p. 12 for the definition.

at most one outgoing edge "leading" to customer classes of lower priority, which follows from the way we assigned priority to customer classes. We shall assume that the server classes are numbered so that the lowest-priority activity is (IJ).

REMARK 2.9. This procedure will not give a unique assignment of priorities: choosing the leaves in different orders will result in different assignments. We give two examples in Figure 2.3. The LAP analysis applies equally well to any priority assignment constructed

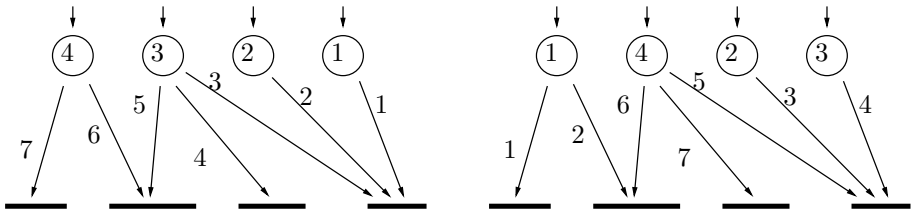

FIGURE 2.3. Two examples of assigning priorities to customer classes and activities of the same tree.

according to the above algorithm.

DEFINITION 2.10. We will write $(ij) < (i'j')$ to mean that activity (ij) has higher priority than activity $(i'j')$.

For example, if $j = j'$, we have $(ij) < (i'j)$ if and only if $i < i'$.

Now we define the LAP policy itself.

Scheduling: A server of type j upon completing a service picks the customer from the queue of type $i \in \mathcal{C}(j)$ such that $i \leq i'$ for all $i' \in \mathcal{S}(i)$ with $Q_{i'} > 0$. If no customer types in $\mathcal{C}(j)$ have queues, the server remains idle.

Routing: An arriving customer of type i picks an unoccupied server in the pool $j \in \mathcal{S}(i)$ such that $(ij) \leq (ij')$ for all $j' \in \mathcal{S}(i)$ with $Z_{j'} < 0$. If no server pools in $\mathcal{S}(i)$ have idle servers, the customer queues.

4. Fluid-scaled convergence for LQFS-LB and LAP

In this section, we consider the behaviour of large systems under the *fluid scaling* $\gamma^r(\cdot) = \frac{1}{r}\Gamma^r(\cdot)$ for all state variables Γ. This is a rather coarse description of the process; later, we will also investigate the behaviour of $r^{-1/2+\epsilon}(\Gamma^r(\cdot) - r\gamma^*)$, for $0 \leq \epsilon < 1/2$ and some appropriately chosen constant γ^*.

We will show that, under the fluid rescaling, the Markov processes describing the system state converge, as $r \to \infty$, in distribution and uniformly on compact sets, to a set of Lipschitz functions satisfying certain *fluid model equations*. We refer to subsequential limits $\lim_{r_k \to \infty} \gamma^{r_k}(\cdot)$ as *fluid limits*, and to Lipschitz functions satisfying appropriate equations as *fluid models*; in these terms, we show that all fluid limits are fluid models. We will then analyse the behaviour of fluid models (which, by the convergence, will be approximately the same for all sufficiently large systems). Analysis of fluid models is a standard technique in the theory of queueing networks; see, for example, Bramson [2006].

In order to show convergence of processes, we would like to formalize the control we have over the arrival processes. We will assume that the arrival and service processes are rescalings of a family of independent unit-rate Poisson processes. Moreover, for any sequence $r \to \infty$ there is a subsequence (also indexed by r) along which the underlying Poisson processes are "well-behaved"; we will work only along such subsequences.

Formally, we make the following assumptions:

ASSUMPTION 2.11. *Let $\Pi^{(a)}_{\mathcal{I}}(\cdot)$ and $\Pi^{(s)}_{\mathcal{E}}(\cdot)$ be independent unit-rate Poisson processes. We will assume that, for each r,*

$$A^r_i(t) = \Pi^{(a)}_i(\lambda_i r t),\ \forall i \in \mathcal{I} \quad S^r_{ij}(t) = \Pi^{(s)}_i(\mu_{ij} r t),\ \forall (ij) \in \mathcal{E}.$$

Poisson processes $\Pi^{(a)}_i(\cdot)$ and $\Pi^{(s)}_{ij}(\cdot)$ satisfy the following property (see [Mandelbaum and Stolyar, 2004, (34)]). Any subsequence of $\{r\}$, has a further subsequence, such that with probability 1, for any fixed $t > 0$ and $d > 0$, uniformly on any sequence of pairs (s^r, t^r) with $0 \le s^r < t^r \le rt$ and $t^r - s^r \ge \sqrt{r}d$, we have

$$\lim_{r\to\infty} \frac{\Pi^{(a)}_i(t^r) - \Pi^{(a)}_i(s^r)}{t^r - s^r} = 1 \tag{4}$$

and similarly for $\Pi^{(s)}_{ij}(\cdot)$. This lets us work with pathwise limits, rather than limits in distribution.

ASSUMPTION 2.12. *The sequence $\{r\}$ is such that (4) holds for $\Pi^{(a)}_i(\cdot)$ and $\Pi^{(s)}_{ij}(\cdot)$, for all $i \in \mathcal{I}$ and $(ij) \in \mathcal{E}$.*

4.1. Convergence for LQFS-LB. Consider the scaling

$$\Big(\psi^r_{\mathcal{E}}(t), q^r_{\mathcal{I}}(t), x^r_{\mathcal{I}}(t), a^r_{\mathcal{I}}(t), \rho^r_{\mathcal{J}}(t)\Big) \equiv \frac{1}{r}\Big(\Psi^r_{\mathcal{E}}(t), Q^r_{\mathcal{I}}(t), X^r_{\mathcal{I}}(t), A^r_{\mathcal{I}}(t), P^r_{\mathcal{J}}(t)\Big)$$

THEOREM 2.13. *Suppose*

$$(\psi^r_{\mathcal{E}}(0), q^r_{\mathcal{I}}(0)) \to (\psi_{\mathcal{E}}(0), q_{\mathcal{I}}(0))\,.$$

Then, w.p.1, for any sequence $r \to \infty$ there exists a subsequence along which

$$\big(\psi^r_{\mathcal{E}}(\cdot), q^r_{\mathcal{I}}(\cdot), x^r_{\mathcal{I}}(\cdot), a^r_{\mathcal{I}}(\cdot), \rho^r_{\mathcal{J}}(\cdot)\big) \to (\psi_{\mathcal{E}}(\cdot), x_{\mathcal{I}}(\cdot), q_{\mathcal{I}}(\cdot), a_{\mathcal{I}}(\cdot), \rho_{\mathcal{J}}(\cdot))$$

uniformly on compact sets, for some set of Lipschitz functions

$$(\psi_{\mathcal{E}}, q_{\mathcal{I}}, x_{\mathcal{I}}, a_{\mathcal{I}}, \rho_{\mathcal{J}})$$

satisfying the fluid model equations (5). (The conditions involving derivatives are to be satisfied whenever the derivatives exist, which is Lebesgue-almost everywhere.)

The LQFS-LB fluid model equations are

$$a_i(t) = \lambda_i t,\ \forall i \in \mathcal{I} \tag{5a}$$

$$x_i(t) = q_i(t) + \sum_j \psi_{ij}(t),\ \forall i \in \mathcal{I} \tag{5b}$$

$$x_i(t) = x_i(0) + a_i(t) - \sum_j \int_0^t \mu_{ij}\psi_{ij}(s)ds,\ \forall i \in \mathcal{I} \tag{5c}$$

$$\rho_j(t) = \frac{1}{\beta_j}\sum_i \psi_{ij}(t),\ \forall j \in \mathcal{J} \tag{5d}$$

$$\rho_j(t) = 1 \text{ if } q_i(t) > 0 \text{ for any } i \in \mathcal{C}(j),\ \forall j \in \mathcal{J} \tag{5e}$$

For any set of customer types $\mathcal{I}_* \subseteq \mathcal{I}$, and any set of server types $\mathcal{J}_* \subseteq \mathcal{J}$ such that (a) $\rho_j(t) < 1$ for all $j \in \mathcal{J}_*$, and (b) $\rho_j(t) < \rho_{j'}(t)$ whenever $j \in \mathcal{J}_*$, $j' \notin \mathcal{J}_*$, and $\mathcal{C}(j) \cap \mathcal{C}(j') \cap \mathcal{I}_* \neq \emptyset$,

$$\sum_{j\in\mathcal{J}_*}\sum_{i\in\mathcal{C}(j)\cap\mathcal{I}_*} \dot{\psi_{ij}}(t) = \sum_{i\in\cup_{j\in\mathcal{J}_*}\mathcal{C}(j)\cap\mathcal{I}_*} \lambda_i - \sum_{j\in\mathcal{J}_*}\sum_{i\in\mathcal{C}(j)\cap\mathcal{I}_*} \mu_{ij}\psi_{ij}(t) \tag{5fa}$$

For any set of server types $\mathcal{J}^* \subseteq \mathcal{J}$, and any set of customer types $\mathcal{I}^* \subseteq \mathcal{I}$ such that (a) $q_i(t) > 0$ for all $i \in \mathcal{I}^*$, and (b) $q_i(t) > q_{i'}(t)$ whenever $i \in \mathcal{I}^*$, $i' \notin \mathcal{I}^*$ and $\mathcal{S}(i) \cap \mathcal{S}(i') \cap \mathcal{J}^* \neq \emptyset$,

$$\text{(5fb)} \qquad \sum_{i \in \mathcal{I}^*} \sum_{j \in \mathcal{S}(i) \cap \mathcal{J}^*} \dot{\psi}_{ij}(t) = \sum_{j \in \cup_{i \in \mathcal{I}^*} \mathcal{S}(i) \cap \mathcal{J}^*} \sum_{i' \in \mathcal{C}(j)} \mu_{i'j} \psi_{i'j}(t) - \sum_{i \in \mathcal{I}^*} \sum_{j \in \mathcal{S}(i) \cap \mathcal{J}^*} \mu_{ij} \psi_{ij}(t)$$

We comment on (5f) as the least intuitive. It describes the idea that customers are only entering service at one of the least busy servers that they can find, while servers are only taking requests from one of the longest queues that they can serve.

The meaning of (5fa) is as follows. Consider a set of customer types $\mathcal{I}_*$. If a set of server types $\mathcal{J}_*$ consists of the "least busy server types for $\mathcal{I}_*$" (we will make this more precise), then arriving customers of type $i_* \in \mathcal{I}_*$ will all be routed to servers in $\mathcal{J}_*$. In this case, the total number of customers of types $\mathcal{I}_*$ in service by servers of types $\mathcal{J}_*$ will be changing at the total arrival rate of customers in $\mathcal{I}^*$, less the rate of servicing customers of all types by servers in $\mathcal{J}^*$.

The requirements that $\mathcal{J}_*$ needs to satisfy for this to be the case are, that there be no server types outside $\mathcal{J}_*$ with smaller instantaneous load, which can serve customers of some type in $\mathcal{I}_*$. We now consider some examples of what valid sets $\mathcal{J}_*$ can look like for a given $\mathcal{I}_*$. As warm-up, a one-element set $\mathcal{J}_* = \{j_*\}$ is a valid choice for a one-element set $\mathcal{I}_* = \{i_*\}$ if and only if the server pool $j_* \in \mathcal{S}(i_*)$ has the (strictly) smallest instantaneous load among all of the server types that can serve i_*.

Consider now the situation in the right-hand network of Figure 2.4. If $\mathcal{I}_* = \{A, B, C, D\}$, then the only valid choices of $\mathcal{J}_*$ are $\{a\}$ and $\{a, b, c, d\}$. Note that $\mathcal{J}_0 \equiv \{a, b\}$ does not qualify, because b shares a customer type $B \in \mathcal{I}_*$ with an equally-loaded server pool $c \notin \mathcal{J}_0$. If we instead look at $\mathcal{I}_* = \{C, D\}$, then $\mathcal{J}_* = \{d\}$ becomes a valid choice: no $j \in \{d\}$ and $j' \in \{a, b, c\}$ "share" a customer type in $\mathcal{I}_*$.

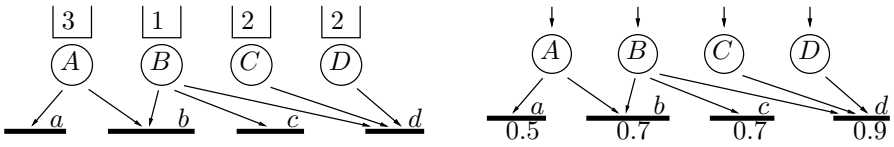

FIGURE 2.4. Illustration for (5f).

In (5fb) we consider a similar situation, but with queueing: in this case, servers are picking customers, and not the other way around. Consider a set of server types $\mathcal{J}^*$. If a set of customer types $\mathcal{I}^*$ consists of the "longest queues for $\mathcal{J}^*$" (we will make this more precise), then servers in pools $j^* \in \mathcal{J}^*$, whenever they finish serving some customer, will immediately replace her with someone from queue i^*. In this case, the total number of customers of types $\mathcal{I}^*$ in service by servers of types $\mathcal{J}^*$ will be increasing at the total rate of servicing all customers by servers in $\mathcal{J}^*$, less the rate of servicing customers of types $\mathcal{I}^*$ by servers in $\mathcal{J}^*$.

The requirements that $\mathcal{I}^*$ needs to satisfy for this to be the case are, that there be no customer types outside $\mathcal{I}^*$ with longer queues that servers in $\mathcal{J}^*$ can serve. We now consider some examples of what valid sets $\mathcal{I}^*$ can look like for a given $\mathcal{J}^*$. As warm-up, a one-element set $\mathcal{I}^* = \{i^*\}$ is a valid choice for a one-element set $\mathcal{J}^* = \{j^*\}$ if and only if the customer type $i^* \in \mathcal{C}(j^*)$ has the (strictly) longest queue among all of the customer types that can be served by j^*.

Consider now the situation in the left-hand network of Figure 2.4. If $\mathcal{J}^* = \{a, b, c, d\}$, then the valid choices of $\mathcal{I}^*$ are $\{A\}$, $\{C, D\}$, and $\{A, B, C, D\}$. Note that $\mathcal{I}^0 \equiv \{C\}$ alone does not qualify, because C shares a server type $d \in \mathcal{J}^*$ with a queue of the same length

$D \notin \mathcal{I}^0$. On the other hand, the fact that $q_A > q_C$ does not stop $\{C, D\}$ from being a valid choice for $\mathcal{I}^*$, because $\mathcal{S}(i) \cap \mathcal{S}(i') = \emptyset$ for $i \in \{A\}$, $i' \in \{C, D\}$. If we instead look at $\mathcal{J}^* = \{c\}$, then $\mathcal{I}^* = \{B, C\}$ becomes a valid choice: no $i \in \{B, C\}$ and $i' \in \{A, D\}$ "share" a server type in $\mathcal{J}^*$.

Proof of Theorem 2.13. Given property (4), it is standard to conclude that, with probability 1, any sequence of fluid-scaled processes has a subsequence which converges uniformly on compact sets to some absolutely continuous[6] limit; see for example [Mandelbaum et al., 1998, Theorem 2.1]. That the limit is then Lipschitz follows from the fact that the arrival rate and the maximal service rate of customers are upper-bounded. We will skip the technical difficulties of demonstrating the existence of Lipschitz limits, and focus instead on the question of why any fluid limit must satisfy the fluid model equations (5f).

(5a) is a direct consequence of (4). (5b) holds in all prelimit systems, hence in the limit as well.

(5c) also follows from (4). Indeed, in the prelimit system we have

$$X_i^r(t) = X_i^r(0) + A_i^r(t) - \sum_{j \in \mathcal{S}(i)} \Pi_{ij}^{(s)} \left(\int_0^t \mu_{ij} \Psi_{ij}^r(s) ds \right).$$

Dividing by r and using the fact that the limit as $r \to \infty$ exists and is Lipschitz, we can apply (4) to conclude

$$\frac{1}{r} \Pi_{ij}^{(s)} \left(\int_0^t \mu_{ij} \Psi_{ij}^r(s) ds \right) \to \int_0^t \mu_{ij} \psi_{ij}(s) ds.$$

(5d) holds in all prelimit systems, hence in the limit as well. (5e) follows from the fact that, in the rth system, $\rho_j^r(t) = 1$ whenever any customer type $i \in \mathcal{C}(j)$ has a positive queue. (Note that if $q_i > 0$ then $q_i^r > 0$ for all sufficiently large r.)

We now turn to (5f). Recall that the limit is Lipschitz, hence absolutely continuous, so the equation makes sense almost everywhere. Let t be one of the regular times at which the derivatives of all of the limiting process $\dot{\psi}_{ij}(t)$ exist.

Consider (5fb). Pick a set of server types $\mathcal{J}^* \subseteq \mathcal{J}$, and a set of customer types $\mathcal{I}^* \subseteq \mathcal{I}$ satisfying the conditions. Since $q_i(t) > q_{i'}(t)$ for all $i \in \mathcal{I}^*$, $i' \notin \mathcal{I}^*$ s.t. $\mathcal{S}(i) \cap \mathcal{S}(i') \cap \mathcal{J}^* \neq \emptyset$, there exists a $\delta > 0$ sufficiently small that $q_i(s) > q_{i'}(s) + \delta$ for all $s \in [t, t + \delta)$; and then for all sufficiently large r we have $q_i^r(s) > q_{i'}^r(s) + \delta/2$ for all $s \in [t, t + \delta)$. Consequently, during the entire time interval $[t, t+\delta]$, all servers in $\mathcal{J}^*$ that can take customers in $\mathcal{I}^*$ will do so. Now, during $[t, t + \delta]$, each server type $j \in \mathcal{J}^*$ has approximately $\sum_{i' \in \mathcal{C}(j)} \mu_{i'j} \Psi_{i'j}^r \delta$ service completions (this is a consequence of (4)), all of which are replaced by customers of types $i \in \mathcal{I}^*$. Therefore, the total fluid-scaled number of customers of types in $\mathcal{I}^*$ being served by servers in $\mathcal{J}^*$ will be changing by approximately

$$\delta \left(\sum_{j \in \cup_{i \in \mathcal{I}^*} \mathcal{S}(i) \cap \mathcal{J}^*} \sum_{i' \in \mathcal{C}(j)} \mu_{i'j} \psi_{i'j}^r(t) - \sum_{i \in \mathcal{I}^*} \sum_{j \in \mathcal{S}(i) \cap \mathcal{J}^*} \mu_{ij} \psi_{ij}^r(t) \right).$$

Since we are assuming all the derivatives $\dot{\psi}_{ij}(t)$ exist, they must satisfy (5fb).

The argument for (5fa) is nearly identical. Picking a regular time t, let δ be small enough that strict inequalities on instantaneous loads in the limiting system hold at all times in $[t, t + \delta]$ in the prelimit fluid-scaled systems for all sufficiently large r. Then all customer arrivals to types in $\mathcal{I}^*$ must be routed to servers in $\mathcal{J}^*$, so the fluid-scaled

[6] See Appendix A for a definition of absolute continuity.

number of customers of types in $\mathcal{I}^*$ being served by servers in $\mathcal{J}^*$ will be changing by approximately

$$\delta \left(\sum_{i \in \cup_{j \in \mathcal{J}_*} \mathcal{C}(j) \cap \mathcal{I}_*} \lambda_i - \sum_{j \in \mathcal{J}_*} \sum_{i \in \mathcal{C}(j) \cap \mathcal{I}_*} \mu_{ij} \psi_{ij}^r(t) \right).$$

□

DEFINITION 2.14. We call any Lipschitz solution of (5)

$$(\psi_{\mathcal{E}}(\cdot), q_{\mathcal{I}}(\cdot), x_{\mathcal{I}}(\cdot), a_{\mathcal{I}}(\cdot), \rho_{\mathcal{J}}(\cdot))$$

a *fluid model* of the LQFS-LB system with initial state $(\psi_{\mathcal{E}}(0), q_{\mathcal{I}}(0))$; a set $(\psi_{\mathcal{E}}(\cdot), q_{\mathcal{I}}(\cdot))$, which is a projection of a fluid model, we often call a fluid model as well.

REMARK 2.15. In general, the set of fluid models will be larger than the set of fluid-scaled limits of queueing processes. Further, in general, given a set of initial conditions, there need not be a unique fluid model starting from that set of conditions; indeed, there may not even be a unique fluid limit. For the rest of the exposition, it will not be important whether solutions to the fluid model equations are uniquely defined by their starting state; but it is an interesting question in its own right. We show below that, indeed, there is a unique fluid limit from any starting state, and in the process derive the additional equations that need to be added to the (5) to enforce this uniqueness.

Consider the quantity $\xi_{ij}(t)$, the amount of customers of type i that have been routed to servers of type j up to time t. (We define $\xi_{ij}(0) = 0$ for concreteness.) It is not hard to see that $\xi_{\mathcal{E}}(\cdot)$ will be Lipschitz, and that knowing the initial state of the system $(\psi_{\mathcal{E}}(0), q_{\mathcal{I}}(0))$ and $\xi_{\mathcal{E}}(\cdot)$ is equivalent to knowing the entire trajectory of the fluid model.

Let $\lambda_{\mathcal{E}}(\cdot) = \frac{d}{dt}\xi_{\mathcal{E}}(t)$ whenever this is defined; $\lambda_{ij}(t)$ is the instantaneous rate of routing customers of type i to servers of type j. Since $\xi_{\mathcal{E}}(t)$ is Lipschitz, $\lambda_{\mathcal{E}}(\cdot)$ determines $\xi_{\mathcal{E}}(\cdot)$. We will now show that, given the state of the fluid model at time t, $\lambda_{\mathcal{E}}(t^+)$ is uniquely determined. (Note that we already know that one feasible $\lambda_{\mathcal{E}}(t^+)$ exists, because the fluid limit started from that initial state will determine some set of values.) We will usually drop the time index t^+ in what follows.

Note firstly that $\lambda_{ij} = 0$ if (a) there exists some $i' \in \mathcal{C}(j)$ with $q_{i'} > q_i$, or (b) there exists some $j' \in \mathcal{S}(i)$ with $\rho_{j'} < \rho_j$. Consequently, we have partitioned the basic activity tree $\mathcal{E}$ into subtrees, such that within each subtree all customer queue sizes and all server pool loads are equal. We will now restrict attention to one such subtree, T; WLOG it will be a subtree with all queue sizes equal to $q > 0$. Let C and S denote the subsets of customer and server types belonging to T.

If we were given the constraint that all queues in T stay equal at t^+, then we could determine the routing rates λ_{ij}. Indeed, if all queue sizes remain equal, then necessarily

$$\dot{q}(t^+) = |C|^{-1} \left(\sum_{i \in C} \lambda_i - \sum_{i' \in \cup_{j \in S} \mathcal{C}(j)} \psi_{i'j}(t) \mu_{i'j} \right),$$

and $\dot{\rho}_j(t) = 0$ for all $j \in C$ (since $q > 0$). This allows us to solve for $\lambda_{ij}(t^+)$ by sequentially eliminating leaves of the tree. If customer type i is a leaf with unique server pool j, then $\lambda_{ij}(t^+) = \dot{q}(t^+) - \lambda_i$; and if server type j is a leaf with unique customer type i, then $\lambda_{ij}(t^+) = \mu_{ij}(t)\psi_{ij}$.

Unfortunately, this may give $\lambda_{ij}(t^+) < 0$ for some activities, which is not physical (the process $\xi_{ij}(t)$ must be nondecreasing). This indicates that, in fact, the queue sizes of customer types in C will not remain equal; rather, our tree T will split into subtrees $T_1, T_2, \ldots, T_n$ with the following properties:

(1) Within each T_k, queue sizes will remain equal, and will change at rate $\dot{q}_k(t^+)$ (positive or negative). WLOG, the indexing is such that $\dot{q}_1 > \dot{q}_2 > \ldots > \dot{q}_k$. [7]
(2) The associated rates $\lambda_{ij}(t^+) \geq 0$ within each subtree.
(3) The rates $\lambda_{ij}(t^+)$ are 0 if i and j belong to different subtrees. This means that if, for a basic activity (ij) we get $i \in T_k$ and $j \in T_{k'}$, then $\dot{q}_k < \dot{q}_{k'}$.

Observe that, once we know the subtrees, the associated rates $\lambda_{ij}(t^+)$ are completely determined.

We now claim that the partition P of T into subtrees satisfying (1)–(3) is unique. (Again, we know one exists because the fluid limit with this initial state must give one.) Indeed, suppose $\tilde{P} \equiv \{\tilde{T}_1, \ldots, \tilde{T}_{\tilde{n}}\}$ is another partition, and WLOG let $\dot{q}_1 \geq \dot{\tilde{q}}_1$. Consider now the queues of types $C_1 \in T_1$. The total amount of service that they are getting in the partition $\tilde{P}$ cannot be greater than in P, since in P they are getting all of the servers available to them. Consequently, at least one of these queues will have a higher time derivative in $\tilde{P}$ than $\dot{q}_1$, and equality is only possible if in $\tilde{P}$ the set of queues C_1 also gets, and shares equally, all the service that it can – i.e., if $C_1 = \tilde{C}_1$. Continuing inductively gives the result.

The argument is similar if we restrict our attention to a subtree where all queues are 0.

Thus, we've shown that, for any state $(\psi_{\mathcal{E}}(t), q_{\mathcal{I}}(t))$ there is a unique set of time derivatives $\lambda_{\mathcal{E}}(t^+) \geq 0$ that are consistent with the fluid model equations. Now, (5) has no equations equivalent to the nonnegativity of $\lambda_{\mathcal{E}}(t^+)$. However, adding this constraint (by adding the process $\xi_{\mathcal{E}}(t)$ to the state descriptor and requiring it to be nondecreasing) would, as we saw above, force uniqueness.

4.2. Convergence for LAP. We now perform similar analysis for the LAP policy. Our state descriptor will need to be slightly larger than for the LQFS-LB model, but otherwise the analysis is very similar.

Consider the scaling

$$\Big(\psi^r_{\mathcal{E}}(t), q^r_{\mathcal{I}}(t), x^r_{\mathcal{I}}(t), a^r_{\mathcal{I}}(t), d^r_{\mathcal{E}}(t), \xi^r_{\mathcal{E}}(t)\Big) \equiv \frac{1}{r}\Big(\Psi^r_{\mathcal{E}}(t), Q^r_{\mathcal{I}}(t), X^r_{\mathcal{I}}(t), A^r_{\mathcal{I}}(t), D^r_{\mathcal{E}}(t), \Xi^r_{\mathcal{E}}(t)\Big).$$

PROPOSITION 2.16. *Suppose*

$$(\psi^r_{\mathcal{E}}(0), q^r_{\mathcal{I}}(0)) \to (\psi_{\mathcal{E}}(0), q_{\mathcal{I}}(0))$$

Then, w.p.1, for any sequence $r \to \infty$ there exists a subsequence along which

$$(\psi^r_{\mathcal{E}}(\cdot), q^r_{\mathcal{I}}(\cdot), x^r_{\mathcal{I}}(\cdot), a^r_{\mathcal{I}}(\cdot), d^r_{\mathcal{E}}(\cdot), \xi^r_{\mathcal{E}}(\cdot)) \to (\psi_{\mathcal{E}}(\cdot), q_{\mathcal{I}}(\cdot), x_{\mathcal{I}}(\cdot), a_{\mathcal{I}}(\cdot), d_{\mathcal{E}}(\cdot), \xi_{\mathcal{E}}(\cdot))$$

uniformly on compact sets, for some set of Lipschitz functions

$$(\psi_{\mathcal{E}}, q_{\mathcal{I}}, x_{\mathcal{I}}, a_{\mathcal{I}}, d_{\mathcal{E}}, \xi_{\mathcal{E}})$$

satisfying the fluid model equations (7). *(The conditions involving derivatives are to be satisfied whenever the derivatives exist, which is Lebesgue-almost everywhere.)*

The LAP fluid model equations are

$$\text{(7a)} \qquad q_i(t) \geq 0,\ \forall i \in \mathcal{I}; \quad \psi_{ij}(t) \geq 0,\ \forall (ij) \in \mathcal{E}; \quad \sum_i \psi_{ij}(t) \leq \beta_j,\ \forall j \in \mathcal{J}$$

$$\text{(7b)} \qquad a_i(t) = \lambda_i t,\ \forall i \in \mathcal{I}; \quad d_{ij}(t) = \int_0^t \mu_{ij}\psi_{ij}(s)ds,\ \forall (ij) \in \mathcal{E}$$

[7] If we require strict inequalities between $\dot{q}_k$ for different k, then strictly speaking T_k might end up disconnected. This makes no difference to the analysis.

$$q_i(t) = q_i(0) + a_i(t) - \sum_j \xi_{ij}(t),\ \forall i \in \mathcal{I} \tag{7c}$$

$$\psi_{ij}(t) = \psi_{ij}(0) + \xi_{ij}(t) - d_{ij}(t),\ \forall i \in \mathcal{I} \tag{7d}$$

$$x_i(t) = q_i(t) + \sum_j \psi_{ij}(t) = x_i(0) + \lambda_i t - \sum_j \int_0^t \mu_{ij}\psi_{ij}(s)ds,\ \forall i \in \mathcal{I} \tag{7e}$$

$$\sum_i \psi_{ij}(t) = \beta_j, \text{ whenever } q_{i'}(t) > 0 \text{ for at least one } i' \in \mathcal{C}(j) \tag{7f}$$

$$\frac{d}{dt}\xi_{ij}(t) = 0, \text{ whenever } q_{i'}(t) > 0 \text{ for at least one } i' \in \mathcal{C}(j),\ i' < i \tag{7g}$$

$$\frac{d}{dt}\xi_{ij}(t) = 0, \text{ whenever } \sum_k \psi_{kj'}(t) < \beta_{j'} \text{for at least one } j' \in \mathcal{S}(i) \text{ with } (ij') < (ij) \tag{7h}$$

$$\frac{d}{dt}\xi_{ij}(t) = \min\left(\lambda_i - \sum_{(ij')<(ij)} \frac{d}{dt}\xi_{ij'}(t), \sum_{i'} \mu_{i'j}\psi_{i'j}(t) - \sum_{(i'j)<(ij)} \frac{d}{dt}\xi_{ij'}(t)\right) \tag{7i}$$
$$\text{whenever } q_{i'}(t) = 0 \text{ for all } i' \in \mathcal{C}(j),\ i' < i, \text{ and}$$
$$\sum_k \psi_{kj'} = \beta_{j'} \text{ for all } j' \in \mathcal{S}(i) \text{ with } (ij') < (ij).$$

Proof. Given property (4), it is standard to conclude that, with probability 1, any sequence of fluid-scaled processes has a subsequence which converges uniformly on compact sets to some absolutely continuous limit; see for example [Mandelbaum et al., 1998, Theorem 2.1]. That the limit is then Lipschitz follows from the fact that the arrival rate and the maximal service rate of customers are bounded above. We will skip the technical difficulties of demonstrating the existence of Lipschitz limits, and focus instead on the question of why any fluid limit must satisfy the fluid model equations (7)

(7a) holds in all prelimit systems, hence in the limit. (7b) is a direct consequence of (4). (7c) and (7d) hold in all prelimit systems, hence in the limit.

(7e) also follows from (4). Indeed, in the pre-limit system we have

$$X_i^r(t) = X_i^r(0) + A_i^r(t) - \sum_{j\in\mathcal{S}(i)} \Pi_{ij}^{(s)}\left(\int_0^t \mu_{ij}\Psi_{ij}^r(s)ds\right).$$

Dividing by r and using the fact that the limit as $r \to \infty$ exists and is Lipschitz, we can apply (4) to conclude

$$\frac{1}{r}\Pi_{ij}^{(s)}\left(\int_0^t \mu_{ij}\Psi_{ij}^r(s)ds\right) \to \int_0^t \mu_{ij}\psi_{ij}(s)ds.$$

(7f)–(7h) hold in all prelimit systems, hence in the limit.

Finally, (7i) follows from (7f)–(7h). □

Definition 2.17. We call any Lipschitz solution

$$(\psi_{\mathcal{E}}(\cdot), q_{\mathcal{I}}(\cdot), x_{\mathcal{I}}(\cdot), a_{\mathcal{I}}(\cdot), d_{\mathcal{E}}(\cdot), \xi_{\mathcal{E}}(\cdot))$$

of (7) a *fluid model* of the LAP system with initial state $(\psi_{\mathcal{E}}(0), q_{\mathcal{I}}(0))$; a set $(\psi_{\mathcal{E}}(\cdot), q_{\mathcal{I}}(\cdot))$, which is a projection of a fluid model, we often call a fluid model as well.

REMARK 2.18. In this case, we also have uniqueness of fluid model solutions given starting state, and rather more simply than for LQFS-LB: since LAP is a simple priority discipline, we can directly determine the quantities $\lambda_{ij}(t^+) \equiv \dot{\xi}_{ij}(t^+)$ in order of decreasing priority.

5. LQFS-LB fluid models near equilibrium

5.1. Linear ODE. In this section, we examine the behaviour of the fluid models for LQFS-LB. Define the *equilibrium point* of the LQFS-LB fluid model as follows.

DEFINITION 2.19. In underload ($\rho < 1$), the *equilibrium point* is the state

$$\psi^*_{ij} \equiv \frac{\lambda_{ij}}{\mu_{ij}},\ \forall (ij) \in \mathcal{E}, \quad q^*_i \equiv 0,\ \forall i \in \mathcal{I}$$

where $\lambda_{\mathcal{E}}$ are the optimal solution to the SPP (1) (unique by Assumption 2.4).

In critical load ($\rho = 1$), an *equilibrium point* is any state with

$$\psi^*_{ij} \equiv \frac{\lambda_{ij}}{\mu_{ij}},\ \forall (ij) \in \mathcal{E}, \quad q^*_i \equiv q^*,\ \forall i \in \mathcal{I}$$

for some constant $q^* \geq 0$. Thus, in critical load the equilibrium point is non-unique, although $\psi^*_{\mathcal{E}}$ is uniquely defined.

It is easy to see that the functions $(\psi_{\mathcal{E}}(t), q_{\mathcal{I}}(t)) = (\psi^*_{ij}, q^*_i)$ for all t are indeed a fluid model.

DEFINITION 2.20. The values associated with the equilibrium point are henceforth referred to as *nominal*. For example, ψ^*_{ij} is the nominal occupancy (of pool j by type i), λ_i is the nominal arrival rate, λ_{ij} is the nominal routing rate (along activity (ij)), $\psi^*_{ij}\mu_{ij} = \lambda_{ij}$ is the nominal service rate (of type i in pool j), $\sum_j \psi^*_{ij}\mu_{ij} = \lambda_i$ is the nominal total service rate (of type i), ρ is the nominal total occupancy (of each pool j), etc.

Desirable system behaviour would be to have $\psi_{\mathcal{E}}(t) \to \psi^*_{\mathcal{E}}$ as $t \to \infty$; we will now investigate whether this in fact occurs.

We will consider two cases: $\rho < 1$ and $(\rho = 1, q^* > 0)$[8]. In the first case, in a sufficiently small neighbourhood of the equilibrium, the system state can be described by specifying the $I + J - 1$ variables $\psi_{ij}(t)$. In the second case, in a sufficiently small neighbourhood of the equilibrium, the system state can also be described by specifying $I + J - 1$ variables, namely, $q_i(t)$ and, for each $j \in \mathcal{J}$, all but one of the $\psi_{ij}(t)$ (for a total of $J - 1$ variables $\psi_{ij}(t)$). Indeed, the condition $q_i(t) > 0$ for all i will imply $\sum_i \psi_{ij}(t) = \beta_j$ for all $j \in \mathcal{J}$. Since the state descriptor has this form on a neighbourhood of the equilibrium point, and the fluid models are Lipschitz, there will be an interval of time during which the state descriptor of the fluid model trajectory is of this form.

We now prove two *state space collapse* results (in underload and in critical load). These results show that, after a finite time, the fluid models are confined to a submanifold of dimension I, rather than $I + J - 1$. In the process, we confirm that LQFS-LB *does* work as a load-balancing mechanism, in that the instantaneous load $\rho_j(t)$ on all of the server types will be equal after a finite time.

THEOREM 2.21. *Let $\rho < 1$. There exists a sufficiently small $\epsilon > 0$, depending only on the system parameters, such that for all sufficiently small δ the following holds. There*

[8]We do not consider here the case $q^* = 0$ in critical load, because for a fluid model it is "atypical": it requires the system workload to be "just right". We will return to the case $\rho = 1$, $q^* = 0$ in §7.4, when we discuss the Halfin-Whitt asymptotic regime.

exist times $T_1 = T_1(\delta)$ and $T_2 = T_2(\delta)$, $0 < T_1 < T_2$, such that if the initial system state $\psi_{\mathcal{E}}(0)$ satisfies

$$\|\psi_{\mathcal{E}}(0) - \psi^*_{\mathcal{E}}\| < \delta,$$

then for all $t \in [T_1, T_2]$ the system state satisfies

$$\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon,$$
$$\rho_j(t) = \rho_{j'}(t) \text{ for all } j, j' \in \mathcal{J}.$$

Moreover, $T_1(\delta) \downarrow 0$ and $T_2(\delta) \uparrow \infty$ as $\delta \downarrow 0$. The evolution of the system during $[T_1, T_2]$ is described by a linear ODE, specified below by (12).

THEOREM 2.22. *Let $\rho = 1$, and consider an equilibrium point with $q^* > 0$. There exists a sufficiently small $\epsilon > 0$, depending only on the system parameters, such that for all sufficiently small $\delta > 0$ the following holds. There exist $T_1 = T_1(\delta)$ and $T_2 = T_2(\delta)$, $0 < T_1 < T_2$, such that if the initial system state satisfies*

$$\|\psi_{\mathcal{E}}(0) - \psi^*_{\mathcal{E}}\| < \delta, \quad \left\|q_{\mathcal{I}}(0) - (q, \ldots, q)^\top\right\| < \delta,$$

then for all $t \in [T_1, T_2]$ the system state satisfies

$$\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon, \quad \left\|q_{\mathcal{I}}(t) - (q, \ldots, q)^\top\right\| < \epsilon,$$
$$q_i(t) = q_{i'}(t) \text{ for all } i, i' \in \mathcal{I}.$$

Moreover, $T_1(\delta) \downarrow 0$ and $T_2(\delta) \uparrow \infty$ as $\delta \downarrow 0$. The evolution of the system during $[T_1, T_2]$ is described by a linear ODE specified below by (14).

PROOF OF THEOREM 2.21. Let us choose a suitably small $\epsilon > 0$. In particular, we require ϵ to be sufficiently small that if $\|\psi_{\mathcal{E}}(0) - \psi^*_{\mathcal{E}}\| < \epsilon$, then $\sum_i \psi_{ij}(t) < \beta_j$ for all j, so there is no queueing. Because the fluid model trajectories are continuous, we can always choose some $T_2 > 0$ such that, for all sufficiently small $\delta > 0$, if $\|\psi_{\mathcal{E}}(0) - \psi^*_{\mathcal{E}}\| < \delta$, then $\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon$ for all $t \le T_2$. We now show that $\rho_j(t) = \rho_{j'}(t)$ for all $j, j' \in \mathcal{J}$, during the time interval $[T_1, T_2]$, for some T_1 depending on δ.

Consider $\rho_*(t) \equiv \min_j \rho_j(t)$, $\rho^*(t) \equiv \max_j \rho_j(t)$, and assume $\rho_*(t) < \rho^*(t)$. Let $\mathcal{J}_*(t) \equiv \{j : \rho_j(t) = \rho_*(t)\}$. Then the total arrival rate to servers of type $j \in \mathcal{J}_*(t)$ is

$$\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*(t)} \lambda_i.$$

We claim that this is strictly greater (by a constant) than the nominal arrival rate $\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*(t)} \lambda_{ij}$. Indeed, under Assumption 2.4 the basic activity tree $\mathcal{E}$ is connected, and $\mathcal{J}_*(t) \subsetneq \mathcal{J}$ (else we couldn't have $\rho_*(t) < \rho^*(t)$). Consequently, we must have $\lambda_{ij'} > 0$ for at least one edge $(ij') \in \mathcal{E}$ such that $i \in \cup_{j \in \mathcal{J}_*(t)} \mathcal{C}(j)$ but $j' \notin \mathcal{J}_*(t)$, i.e.

$$\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}^*(t)} \lambda_i - \sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*(t)} \lambda_{ij} \ge \lambda_{ij'} > 0.$$

Taking the minimum of the $\lambda_{ij'}$ over all (nonempty, proper) subsets $\mathcal{J}_*(t) \subsetneq \mathcal{J}$ gives

$$\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}^*(t)} \lambda_i - \sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*(t)} \lambda_{ij} \ge c > 0$$

for some constant c which depends only on the solution to the static planning problem (1), i.e. only on the system parameters. This inequality holds at all times t such that $\mathcal{J}_*(t) \neq \mathcal{J}$, i.e. $\rho_*(t) < \rho^*(t)$.

On the other hand, the total rate at which customers depart from servers in $\mathcal{J}_*(t)$ is

$$\sum_{i \in \mathcal{C}(j), j \in \mathcal{J}_*(t)} \mu_{ij} \psi_{ij}(t),$$

which for $\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon$ is close to nominal. If we choose $\epsilon < c/2$, we see that arrivals exceed services by at least a constant at all times t such that $\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon$ and $\rho_*(t) < \rho^*(t)$. Similarly (decreasing ϵ if necessary), $\rho^*(t)$ is decreasing at a rate bounded below by a (possibly different) constant. We conclude that while $\rho^*(t) - \rho_*(t) > 0$ (and $\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon$ continues to hold), the difference $\rho^*(t) - \rho_*(t)$ is decreasing at a rate bounded below by a constant. This difference is bounded below by 0, and, being Lipschitz, it is equal to the integral of its own derivative (see Appendix A). Consequently, in finite time $T_1 = T_1(\delta)$, we must reach the set $\rho_*(t) = \rho^*(t)$. Of course, this requires $T_1(\delta) < T_2$, but since T_1 is linear in δ, we can always choose δ small enough for this to hold. Moreover, we clearly have $T_1(\delta) \downarrow 0$ as $\delta \downarrow 0$. Since, as we saw above, the derivative of $\rho^*(t) - \rho_*(t)$ is negative whenever $\rho^*(t) - \rho_*(t) > 0$, and the function $\rho^*(\cdot) - \rho_*(\cdot)$ is equal to the integral of its derivative, the equality $\rho^*(t) = \rho_*(t)$ will continue to hold for $T_1 \leq t \leq T_2$.

It remains to derive the differential equation, and to show that T_2 can be chosen depending on δ so that $T_2 \uparrow \infty$ as $\delta \downarrow 0$.

Once we are confined to the manifold $\rho_j(t) = \rho_{j'}(t) = \rho(t)$ for all t, the system evolution is determined in terms of only I independent variables. Recall that there is no queueing for $t \leq T_2$, so we can take the I variables to be $\psi_i(t)$. Given $\psi_{\mathcal{I}}(t)$, we know $\rho(t)$ as $(\sum_i \psi_i(t))/(\sum_j \beta_j)$. Consequently, we know $\sum_i \psi_{ij}(t) = \rho(t)\beta_j$ and $\sum_j \psi_{ij}(t) = \psi_i(t)$. On a tree, this allows us to solve for $\psi_{ij}(t)$ by "stripping off" leaves. (For a customer type leaf, $\psi_{ij}(t) = \psi_i(t)$, while for a server-type leaf, $\psi_{ij}(t) = \rho(t)\beta_j$; see (15) below.) The resulting relationship will clearly be linear, i.e.

$$(\psi_{ij}(t)) \equiv M(\psi_i(t)) \tag{8}$$

for some matrix M. For future reference, we define the ("load balancing") linear mapping M from $y \in \mathbb{R}^I$ to $z = z_{\mathcal{E}} \in \mathbb{R}^{I+J-1}$ as follows: $z = My$ is the unique solution of

$$\eta = \frac{\sum_i y_i}{\sum_j \beta_j};\ \sum_i z_{ij} = \eta\beta_j, \forall j;\ \sum_j z_{ij} = y_i, \forall i. \tag{9}$$

Let $\mathcal{M}$ denote the manifold containing the image of M; that is,

$$\mathcal{M} \equiv \{My, y \in \mathbb{R}^I\} \subseteq \mathbb{R}^{I+J-1}. \tag{10}$$

Thus, the assertion that $\rho_j(t) = \rho_{j'}(t) = \rho(t)$ for all t is equivalent to the assertion that $\psi_{\mathcal{E}}(t) \in \mathcal{M}$.

The evolution of $\psi_i(t)$ is given by

$$\dot{\psi}_i(t) = \lambda_i - \sum_j \mu_{ij}\psi_{ij}(t),\ \forall i. \tag{11}$$

(This is (5c) for the case of $q_i = 0$, i.e. $x_i(t) = \psi_i(t)$.) By the above arguments we see that this entails (in matrix form)

$$\dot{\psi_{\mathcal{I}}}(t) = \lambda_{\mathcal{I}} + A_u\psi_{\mathcal{I}}(t), \tag{12}$$

where A_u is an $I \times I$ matrix,

$$A_u = SDM; \tag{13}$$

here, M is given by (9), D is the diagonal matrix of service rates with entries $\mu_{\mathcal{E}}$, and S has entries $S_{i,(kj)} = -\delta_{ik}$.

It remains to justify the claim that $T_2(\delta) \uparrow \infty$ as $\delta \downarrow 0$. This follows from the fact that, as long as $t \geq T_1$ and $\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon$, the evolution of the system is described by the linear ODE (12). The solutions have the general form

$$\psi_{\mathcal{I}}(t) - \psi^*_{\mathcal{I}} = \exp(A_u(t - T_1))(\psi_{\mathcal{I}}(T_1) - \psi^*_{\mathcal{I}}), \quad \psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}} = M(\psi_{\mathcal{I}}(t) - \psi^*_{\mathcal{I}})$$

where M and A_u are constant matrices depending on the system parameters. Therefore, if $\|\psi_{\mathcal{I}}(T_1) - \psi^*_{\mathcal{I}}\| \leq \delta$ is sufficiently small, then the time it takes for $\psi_{\mathcal{E}}(t)$ to escape the set $\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon$ can be made arbitrarily large. Since as $\delta \downarrow 0$ we have $T_1(\delta) \downarrow 0$, and the system is Lipschitz, taking $\|\psi_{\mathcal{E}}(0) - \psi^*_{\mathcal{E}}\| < \delta$ for small enough δ will guarantee that $\|\psi_{\mathcal{I}}(T_1) - \psi^*_{\mathcal{I}}\|$ is small, and hence we can choose $T_2(\delta) \uparrow \infty$. □

The proof of Theorem 2.22 proceeds similarly; we outline only the differences.

Proof of Theorem 2.22. We take T_2 s.t. $\left\|q_{\mathcal{I}} - (q, \ldots, q)^\top\right\| < \epsilon$ for all $t \leq T_2$. We will take $\epsilon > 0$ sufficiently small that this implies, in particular, $q_i(t) > 0$ for all $i \in \mathcal{I}$, and hence $\rho_j(t) = 1$ for all $j \in \mathcal{J}$, at all times $t \leq T_2$. The equality of queue lengths in $[T_1, T_2]$ is shown analogously to the proof of $\rho_*(t) = \rho^*(t)$ for the underloaded case. Namely, the smallest queue must increase and the largest queue must decrease (as long as not all $q_i(t)$ are equal), because it is getting less (respectively more) service than nominal (we choose ϵ small enough for this to be true provided $\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| < \epsilon$). Thus, in $[T_1, T_2]$ we will have $q_i(t) = q_{i'}(t)$ for all $i, i' \in \mathcal{I}$.

The linear equation is modified as follows. We have

$$\dot{x_i}(t) = \lambda_i - \sum_j \mu_{ij}\psi_{ij}(t).$$

Since we know that all $q_i(t)$ are equal and positive, we have $q_i(t) = q(t) = \frac{1}{I}(\sum x_k(t) - \sum \beta_j)$, and therefore

$$\dot{\psi_i}(t) = \dot{x_i}(t) - \frac{1}{I}\sum_k \dot{x_k}(t).$$

The rest of the argument proceeds as above to give

$$(\dot{\psi_i}(t)) = (\lambda_i - \frac{1}{I}\sum_i \lambda_i) + A_c(\psi_i(t)) \tag{14}$$

for the appropriate matrix A_c which can be computed explicitly from the basic activity tree. The trajectory $\psi_{\mathcal{I}}(\cdot)$ determines $\psi_{\mathcal{E}}(\cdot)$ on $[T_1, T_2]$ (because we are load-balanced with $\rho = 1$), and this in turn determines $x_{\mathcal{I}}(\cdot)$ and $q_{\mathcal{I}}(\cdot)$.

Just as above, the existence of the linear ODE, together with the fact that $T_1(\delta) \downarrow 0$ as $\delta \downarrow 0$, implies that $T_2(\delta) \uparrow \infty$ as $\delta \downarrow 0$. □

To compute the matrix entries of M of (9), and then of A_u, A_c, we carry out the "leaf-stripping" procedure mentioned in the proof of Theorem 2.21. We arrive at the following formula:

$$\psi_{i_0 j_0}(t) = \sum_{i \preceq (i_0, j_0)} \psi_i(t) - \sum_{j \preceq (i_0, j_0)} \rho(t)\beta_j = \tag{15}$$

$$\frac{1}{\sum \beta_j}\left(\sum_{i \preceq (i_0,j_0)} \sum_{j \preceq (j_0,i_0)} \psi_i(t)\beta_j - \sum_{i \preceq (j_0,i_0)} \sum_{j \preceq (i_0,j_0)} \psi_i(t)\beta_j\right)$$

Here, the relation $\preceq$ is defined as follows. Suppose we disconnect the basic activity tree by removing the edge (i_0, j_0). Then for any node k (either customer type or server type) we say $k \preceq (i_0, j_0)$ if it falls in the same component as i_0; otherwise, $k \preceq (j_0, i_0)$. (This is unrelated to the use of $\prec$ for partial orderings in Chapter 3.)

Since in underload we have

$$\dot{\psi_i}(t) = \lambda_i - \sum_j \mu_{ij}\psi_{ij}(t),$$

we obtain the following expression for A_u:

LEMMA 2.23. (1) *Let $\rho < 1$. The entries $(A_u)_{ii'}$ of the matrix A_u are as follows. The coefficient of ψ_i in $\dot{\psi}_i$ is*

$$(A_u)_{ii} = -\frac{1}{\sum_j \beta_j} \sum_{j \in \mathcal{S}(i)} \mu_{ij} \sum_{j' \preceq (j,i)} \beta_{j'}.$$

The coefficient of $\psi_{i'}$ in $\dot{\psi}_i$ is

$$\begin{aligned}(A_u)_{ii'} &= \frac{1}{\sum_j \beta_j} \left[- \sum_{j \in \mathcal{S}(i), j \neq j_{ii'}} \mu_{ij} \sum_{j' \preceq (j,i)} \beta_{j'} + \mu_{ij_{ii'}} \sum_{j' \preceq (i, j_{ii'})} \beta_{j'} \right] \\ &= (A_u)_{ii} + \mu_{ij_{ii'}}.\end{aligned}$$

Here, $j_{ii'} \in \mathcal{S}(i)$ is the neighbour of i such that, after removing the edge $(i, j_{ii'})$ from the basic activity tree, nodes i and i' will be in different connected components. (Such a node is unique, since there is a unique path along the tree from i to i'.)

(2) *The matrix A_u is non-singular.*

(3) *The matrix A_u depends only on $\beta_{\mathcal{J}}$, the basic activity tree structure $\mathcal{E}$, and $\mu_{\mathcal{E}}$, and does* not *depend on $\lambda_{\mathcal{I}}$ and $\psi^*_{\mathcal{E}}$.*

PROOF. (1) We simply use (15) in the expression

$$\frac{d}{dt}\psi_i(t) = \lambda_i - \sum_{j \in \mathcal{S}(i)} \mu_{ij} \psi_{ij}(t).$$

For example, for the network in Figure 2.5, we have

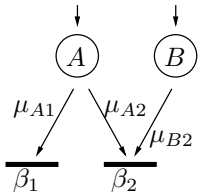

FIGURE 2.5. Example for calculation of the matrix A_u.

$$\begin{pmatrix} \psi_{A1} \\ \psi_{A2} \\ \psi_{B2} \end{pmatrix} = \begin{pmatrix} \frac{\beta_1}{\beta_1+\beta_2} & \frac{\beta_1}{\beta_1+\beta_2} \\ 1 - \frac{\beta_1}{\beta_1+\beta_2} & -\frac{\beta_1}{\beta_1+\beta_2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \psi_A \\ \psi_B \end{pmatrix}$$

giving

$$\begin{pmatrix} \dot{\psi}_A \\ \dot{\psi}_B \end{pmatrix} = \begin{pmatrix} \lambda_A \\ \lambda_B \end{pmatrix} + \begin{pmatrix} -\mu_{A1}\frac{\beta_1}{\beta_1+\beta_2} - \mu_{A2}\frac{\beta_2}{\beta_1+\beta_2} & -\mu_{A1}\frac{\beta_1}{\beta_1+\beta_2} + \mu_{A2}\frac{\beta_1}{\beta_1+\beta_2} \\ 0 & -\mu_{B2} \end{pmatrix} \begin{pmatrix} \psi_A \\ \psi_B \end{pmatrix}$$

as required.

The equality between the two expressions for $(A_u)_{ii'}$ is a consequence of the identity

$$\frac{1}{\sum_j \beta_j} \sum_{j' \preceq (j,i)} \beta_{j'} + \frac{1}{\sum_j \beta_j} \sum_{j' \preceq (i,j)} \beta_{j'} = 1;$$

e.g., in the example above, we observe that

$$-\mu_{A1}\frac{\beta_1}{\beta_1+\beta_2} + \mu_{A2}\frac{\beta_2}{\beta_1+\beta_2} = \left(-\mu_{A1}\frac{\beta_1}{\beta_1+\beta_2} - \mu_{A2}\frac{\beta_2}{\beta_1+\beta_2}\right) + \mu_{A2}.$$

(2) By (12), to show that A_u is nonsingular, it suffices to demonstrate the following. Given a vector of derivatives $\dot{\psi}_{\mathcal{I}}$, we can find a load-balancing vector $\psi_{\mathcal{E}}$ corresponding to some load ρ (an unknown), which results (under (11)) in these derivatives.

Consider the $I+J$ linear equations $\lambda_i - \sum_j \mu_{ij}\psi_{ij} = \dot{\psi}_i$ (for all i) and $\sum_i \psi_{ij} = \rho\beta_j$ (for all j). The value ρ is uniquely determined by the workload derivative condition (see (3)):

$$\sum_i \nu_i \dot{\psi}_i = \sum_i \nu_i \lambda_i - \rho \sum_j \alpha_j.$$

Given the values $\dot{\psi}_{\mathcal{I}}$ and ρ, we can now solve for $\psi_{\mathcal{E}}$ by sequentially eliminating the leaves of the basic activity tree.

(3) Follows from (1).

□

There is also an explicit expression for the entries of A_c, which is obtained similarly:

LEMMA 2.24. (1) *The entries $(A_c)_{ii'}$ of the matrix A_c (for the critical load case, $\rho = 1$) are as follows:*

$$(A_c)_{ii'} = (A_u)_{ii'} - \frac{1}{I}\sum_k (A_u)_{ki'}. \tag{16}$$

(2) *The matrix A_c has rank $I-1$. The $(I-1)$-dimensional subspace $\mathcal{N} = \{y : \sum_i y_i = 0\}$ is invariant under the transformation A_c, i.e. A_c maps vectors in $\mathcal{N}$ to $\mathcal{N}$. Letting π denote the orthogonal projection (along $(1,\ldots,1)^\top$) onto $\mathcal{N}$, we have*

$$A_c = \pi A_u. \tag{17}$$

Restricted to $\mathcal{N}$, the transformation A_c is invertible.

(3) *The linear transformation A_c, restricted to subspace $\mathcal{N}$, depends only on the basic activity tree structure $\mathcal{E}$ and the values $\mu_{\mathcal{E}}$, and does* not *depend on $\beta_{\mathcal{J}}$, $\lambda_{\mathcal{I}}$ and $\psi^*_{\mathcal{E}}$.*

PROOF. (1) The fluid model here is such that there are always non-zero queues, which are equal across customer types. We can write

$$\dot{\psi}_i(t) = \dot{x}_i(t) - \frac{1}{I}\sum_k \dot{x}_k(t) = (\lambda_i - \sum_j \mu_{ij}\psi_{ij}(t)) - \frac{1}{I}\sum_k (\lambda_k - \sum_j \mu_{kj}\psi_{kj}(t)), \tag{18}$$

which implies (16).

(2) First of all, it is not surprising that A_c does not have full rank: the linear ODE defining A_c is such that $\sum_i \psi_i(t) = \sum_j \beta_j$ at all times, so there are at most $(I-1)$ degrees of freedom in the system. Also, it will be readily seen that (16) asserts precisely that $A_c = \pi A_u$. Since A_u is invertible and π has rank $I-1$, their composition has rank $I-1$. Since the image of A_c is contained in $\mathcal{N}$, we must have equality.

It remains to check that A_c restricted to $\mathcal{N}$ still has rank $I-1$. To see this, we observe that the simple eigenvalue 0 of A_c has as its unique right eigenvector the vector $A_u^{-1}(1,1,\ldots,1)^\top$. We will be done once we show that this eigenvector does not belong to $\mathcal{N}$. Suppose instead that $A_u v = (1,1,\ldots,1)^\top$ for some $v \in \mathcal{N}$, $\sum_i v_i = 0$. Then, for small $\epsilon > 0$, starting from some state $\psi_{\mathcal{I}}(t)$, the state $\psi_{\mathcal{I}}(t) + \epsilon v$ would (under balanced loads) have strictly faster service of all the customer types, while keeping the same proportion of servers occupied. This,

however, is impossible. When loads on all the server pools are balanced, the rate at which the system processes workload depends only on the total proportion of occupied servers, hence only on the total number of customers in service.

(3) The specific expression (16) for A_c may depend on the pool sizes $\beta_{\mathcal{J}}$. However, A_c is a singular $I \times I$ matrix, and the statement (3) is only concerned with the transformation of the $(I-1)$-dimensional subspace $\mathcal{N}$ that A_c induces; this transformation does *not* depend on $(\beta_{\mathcal{J}})$, as the following argument shows.

Pick any $(ij) \in \mathcal{E}$. Modify the original system by replacing β_j by $\beta_j + \delta$ and λ_i by $\lambda_i + \delta\mu_{ij}$. Then the ODE (18) for the modified system remains exactly the same as for the original one. Thus, the transformation A_c must not depend on $\beta_{\mathcal{J}}$.

An alternative argument is purely analytic. Recall that to compute $(A_u)_{ij}$ we used (15). In critical load, we have $\rho(t) \equiv 1$, so the (left) equation (15) for $\psi_{i_0 j_0}(t)$ simplifies to

$$\psi_{i_0 j_0}(t) = \sum_{i \preceq (i_0, j_0)} \psi_i(t) - \sum_{j \preceq (i_0, j_0)} \beta_j. \tag{19}$$

If we substitute this in the right-hand side of (18), we will obtain a different expression for $\dot{\psi}_i(t)$. While its constant term will depend on $\beta_{\mathcal{J}}$, the linear term will not, since the linear term of (19) does not depend on $\beta_{\mathcal{J}}$. Therefore, the ODE describing the evolution of $(\psi_{\mathcal{I}} - \psi^*_{\mathcal{I}})$ (which drops the constant term) will not depend on $\beta_{\mathcal{J}}$.

□

We will now analyse the stability of the fluid models for LQFS-LB.

5.2. Definitions of stability.

Definition 2.25. We say that the (fluid) system is *locally stable* if all fluid models starting in a sufficiently small neighborhood of an equilibrium point (which is unique for $\rho < 1$; and for $\rho = 1$ we consider any equilibrium point with equal, positive queues $q^* > 0$) are such that, for some constant $C > 0$ that does not depend on the initial state,

$$\|\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}\| \leq \Delta_0 e^{-Ct},$$

where $\Delta_0 = \|\psi_{\mathcal{E}}(0) - \psi^*_{\mathcal{E}}\| + \left\|q_{\mathcal{I}}(0) - (q^*, \ldots, q^*)^\top\right\|$.

We call the system *globally stable* if any fluid model, with arbitrary initial state, converges to some equilibrium point as $t \to \infty$.

It is not obvious that, as defined here, global stability implies local stability; however, in the example in which we can prove global stability (Theorem 2.30), we shall see that local stability also holds. The assumption of exponential convergence for local stability is the result of Theorems 2.21 and 2.22. The theorems assert that on a neighbourhood of equilibrium the fluid models are governed by a linear ODE, so if they converge at all, they do so exponentially quickly.

Remark 2.26. The definition of global stability implies $\rho_j(t) \to \rho$ for all $j \in \mathcal{J}$ and $\psi_{\mathcal{E}}(t) \to \psi^*_{\mathcal{E}}$ for all $i \in \mathcal{I}$, $j \in \mathcal{J}$. In underload, the definition necessarily implies $q_i(t) \to 0$ for all $i \in \mathcal{I}$. In the case $\rho = 1$, the local stability criterion does not require that $q_i(t) \to q^*$, for q^* associated with the chosen equilibrium point. However, local stability will guarantee convergence of queues *somewhere*. First, if $\psi_{\mathcal{E}}(t) \approx \psi^*_{\mathcal{E}}$ at all times t, then we cannot have large inequalities between queue sizes $q_i(t)$ across different customer types, because the rates at which customers of different type enter service must be approximately nominal. Second, if $\psi_{\mathcal{E}}(t) \approx \psi^*_{\mathcal{E}}$, then system workload is approximately

constant; since the number of customers in service is approximately constant, we conclude that the queues are approximately constant as well. Consequently, local stability will imply that all $q_i(t) \to q$ for some q. In fact, it is not hard to see that $|q - q^*| \leq C_0\Delta_0$ for some constant $C_0 > 0$ depending only on the system parameters. In other words, local stability guarantees convergence to an equilibrium point not too far from the "original" one.

REMARK 2.27. "Global stability" is slightly weaker than the definition of "stability" usually adopted for fluid models. Typically (e.g. [Bramson, 2006, Chapter 4]), a fluid model is called stable if, for all starting states within a ball of radius 1 from the equilibrium point, the fluid model reaches the equilibrium point after a finite time. We, on the other hand, allow convergence to be asymptotic, and do not require uniformity (although in the case of Theorem 2.30 the convergence will indeed be uniform). There is a general theory of proving positive (Harris) recurrence for queueing networks via the stability (in the sense of uniform, finite-time convergence to equilibrium) of fluid networks; see e.g. [Bramson, 2006, Chapter 4]. However, we are not trying to prove stability of LQFS-LB in this sense. In our set-up positive Harris recurrence will hold for all parameter values with $\rho < 1$, because if the queues grow large enough, then all the servers will become fully occupied, and the system will process workload faster than it arrives. (In particular, discussions of "steady state" of the LQFS-LB algorithm are well-defined.) We are interested in the finer question of whether, in steady-state, the LQFS-LB algorithm will eliminate customer queueing, and our notion of global stability is more appropriate.

By Theorems 2.21 and 2.22, local stability is determined by the stability of a linear ODE, which in turn is governed by the eigenvalues of the matrix A_u or A_c.

DEFINITION 2.28. We will call matrix A_u *stable* if all its eigenvalues have negative real part. We call matrix A_c *stable* if all its eigenvalues have negative real part, except for one simple eigenvalue 0[9].

If $A_{u,c}$ is stable, then the corresponding linear ODE (12) or (14) is stable as well. On the other hand, if $A_{u,c}$ has an eigenvalue with positive real part, the ODE has solutions diverging from equilibrium $\psi^*_{\mathcal{I}}$ exponentially fast; and if $A_{u,c}$ has (a pair of conjugate) pure imaginary eigenvalues, the ODE has oscillating, never converging solutions. That is,

PROPOSITION 2.29. *The local stability of the underloaded, respectively critically loaded fluid system is equivalent to the stability of the matrix A_u, respectively A_c.*

We will now examine examples where the matrices $A_{u,c}$ are globally stable, locally stable, and locally unstable. In §5 we will investigate the first and last of these cases further, under the diffusion scaling.

5.3. Global stability. If the service rates in the system depend only on the server type, we have both global and local stability.

THEOREM 2.30. *Assume $\mu_{ij} = \mu_j$ for all $(ij) \in \mathcal{E}$. Then the system is globally stable both for $\rho < 1$ and for $\rho = 1$. In addition, the system is locally stable (i.e. the matrices A_u and A_c are stable).*

[9]A matrix A with all eigenvalues having negative real part is usually called *Hurwitz*. So, A_u stability is equivalent to A_u being Hurwitz; while A_c stability definition is slightly different, because A_c considered as a linear transformation of $\mathbb{R}^I$ is singular. A *symmetric* matrix A, whose eigenvalues are all real, is Hurwitz if and only if it is negative definite, which is a property that can be easily verified by computing some polynomials in the matrix entries (see e.g. [Meyer, 2000, Section 7.6] or [Horn and Johnson, 1985, Section 7.2]). Unfortunately, neither A_u nor A_c is, in general, symmetric; so there appears to be no easy way of determining the sign of the real part of the eigenvalues.

PROOF. Consider the underloaded system, $\rho < 1$, first. First, we show that the lowest load cannot stay too low. Suppose the minimal load $\rho_*(t) \equiv \min_j \rho_j(t)$ is smaller than ρ, and let $\mathcal{J}_*(t) \equiv \{j : \rho_j(t) = \rho_*(t)\}$. Then all customer types in $\mathcal{C}(\mathcal{J}_*(t)) \equiv \bigcup_{j \in \mathcal{J}_*(t)} \mathcal{C}(j)$ are routed to server pools in $\mathcal{J}_*(t)$, so the total arrival rate "into" $\mathcal{J}_*(t)$ is no less than nominal; on the other hand, since $\mu_{ij} = \mu_j$ and server occupancy is lower than nominal, the total departure rate "from" $\mathcal{J}_*(t)$ is smaller than nominal. This shows that if $\rho_* < \rho - \epsilon < \rho$, then $\dot{\rho}_* > \delta > 0$, where $\delta \geq c\epsilon$ for some constant $c > 0$ (depending on the system parameters). That is, if $\rho_*(t) < \rho$ then $\dot{\rho}_*(t) \geq c(\rho - \rho_*(t))$, so $\rho_*(t)$ is bounded below by a function converging exponentially fast to ρ.

Consider a fixed, sufficiently small $\epsilon > 0$; we know that there exists a finite time T_1 such that $\rho_*(t) \geq \rho - \epsilon$ for all $t \geq T_1$. If some customer class i has a queue $q_i(t) > 0$, then all server classes $j \in \mathcal{S}(i)$ have $\rho_j = 1$. It is now easy to see that the system is serving customers faster than they arrive (because $\rho < 1$ and ϵ is small). This easily implies that all $q_i(t) = 0$ after some finite time T_2.

In the absence of queues, we can analyse $\rho^*(t) \equiv \max_j \rho_j(t)$ similarly to the way we treated $\rho_*(t)$; namely, if $\rho^*(t) > \rho$ at some point, then the servers in $\mathcal{J}^*(t) \equiv \{j : \rho_j(t) = \rho^*(t)\}$ are processing workload faster than the nominal rate, and are getting no more arrivals than the nominal quantity. Consequently, $\rho^*(t)$ is bounded above by a function converging exponentially fast to ρ. Since $\rho_*(t) \to \rho$ and $\rho^*(t) \to \rho$, we conclude $\rho_j(t) \to \rho$ for all j.

Once all $\rho_j(t)$ are close enough to ρ, we can use an argument similar to the proof of Theorem 2.21 to conclude that, after a further finite time, we will have $\rho_j(t) = \rho_{j'}(t)$ for all j, j'. (Theorem 2.21 does not apply directly, because we have $\rho_j(t) \approx \rho$, but possibly $\psi_{\mathcal{E}}(t) \not\approx \psi_{\mathcal{E}}$. However, because the service rates $\mu_{ij} = \mu_j$ do not depend on the customer class, we only need the total occupancies of each server pool to be approximately nominal.) Moreover, this common load $\rho(t)$ will satisfy

$$\dot{\rho}(t) \sum_j \beta_j = \sum_i \lambda_i - \rho(t) \sum_j \beta_j \mu_j,$$

and therefore will be given by

$$\rho(t) = \rho + c_1 \exp(-c_2 t)$$

for some constants c_1, $c_2 = (\sum_j \beta_j \mu_j)/(\sum_j \beta_j) > 0$. We conclude that $\rho(t) \to \rho$ (exponentially quickly) and $\dot{\rho}(t) \to 0$ (exponentially quickly).

Define $\hat{\lambda}_{\mathcal{E}}(t)$ by

$$\hat{\lambda}_{ij}(t) \equiv \mu_j \psi_{ij}(t) + \dot{\psi}_{ij}(t). \tag{20}$$

This is the instantaneous rate at which customers of type i are being routed to servers of type j in the absence of queueing. We have $\sum_j \hat{\lambda}_{ij}(t) = \lambda_i$ at all (large) times t. Further, from the discussion of $\rho(t)$ above we conclude

$$\sum_i \hat{\lambda}_{ij}(t) = \mu_j \sum_i \psi_{ij}(t) + \sum_i \dot{\psi}_{ij}(t) = \mu_j \beta_j \rho_j(t) + \beta_j \dot{\rho}_j(t) \to \beta_j \mu_j \rho = \sum_i \lambda_{ij}.$$

This implies $\hat{\lambda}_{\mathcal{E}}(t) \to \lambda_{\mathcal{E}}$, and therefore by (20) $\psi_{\mathcal{E}}(t) \to \psi^*_{\mathcal{E}}$ as required.

Now, consider a critically loaded system, $\rho = 1$. Essentially the same argument as above tells us that, as long as not all queues $q_i(t)$ are equal, each of the longest queues gets more service than the arrival rate into it, and so $q^*(t) \equiv \max q_i(t)$ has derivative which is strictly negative and bounded away from 0. If at some time t, all $q_i(t)$ are equal and positive, then $\dot{q}^*(t) = 0$. We see that $q^*(t)$ is non-increasing, and so $q^*(t) \downarrow q \geq 0$. We

also have $\rho_*(t) \to \rho = 1$ exponentially fast. (Same proof as above applies.) These facts easily imply convergence to an equilibrium point. We omit further detail.

In order to show local stability, it is sufficient to observe that, for all $\epsilon > 0$ there exists a $\delta > 0$ such that fluid models started in a δ-neighbourhood of the equilibrium point will never leave an ϵ-neighbourhood of it. In this case, taking ϵ to be small enough that the behaviour of the fluid model is controlled by a linear ODE ((11) or (14)), convergence to the equilibrium point will imply stability of A_u and A_c. □

5.4. Local stability. If the service rates in the system depend only on the customer type, we have local stability.

THEOREM 2.31. *Assume $\mu_{ij} = \mu_i$ for all $(ij) \in \mathcal{E}$. Then the system is locally stable (i.e. the matrices A_u and A_c are stable).*

PROOF. For the case $\rho < 1$ and $\mu_{ij} = \mu_i$, (11) becomes

$$\dot{\psi}_i(t) = \lambda_i - \mu_i \psi_i(t)$$

and A_u is simply a diagonal matrix with entries $-\mu_i$, which is clearly stable.

Assume now that $\rho = 1$. As we just saw, the matrix A_u in this case is diagonal with entries $-\mu_i$. By Lemma 2.24, A_c has off-diagonal entries $(A_c)_{ii'} = \mu_{i'}/I$ and diagonal entries $-\mu_i(1-1/I)$. In particular, its off-diagonal entries are strictly positive. Therefore, $A_c + \eta I$ for some large enough constant $\eta > 0$ (where I is the identity matrix) is a positive matrix. By Perron-Frobenius theorem [Meyer, 2000, Chapter 8], $A_c + \eta I$ has a real eigenvalue $p + \eta$ with the property that any other eigenvalue of $A_c + \eta I$ is smaller than $p + \eta$ in absolute value (and in particular has real part smaller than $p + \eta$). Moreover, the associated *left* eigenvector w is strictly positive, and is the unique (up to scaling) strictly positive left eigenvector of $A_c + \eta I$. Translating these statements to A_c, we get: A_c has a real eigenvalue p; all other eigenvalues of A_c have real part smaller than p; A_c has a unique (up to scaling) strictly positive left eigenvector w; and the eigenvalue of w is p.

Now, we know that A_c has a positive left eigenvector with eigenvalue 0, namely $(1, 1, \ldots, 1)$. We conclude that $p = 0$, and all other (i.e., non-zero) eigenvalues of A_c have real part smaller than 0, as required. □

5.5. (Local) instability. We have shown that the matrices A_u and A_c are stable in the cases $\mu_{ij} = \mu_j$, $(ij) \in \mathcal{E}$ and $\mu_{ij} = \mu_i$, $(ij) \in \mathcal{E}$. Since the entries of A_u, A_c depend continuously on the parameters $\mu_{\mathcal{E}}$ (Lemmas 2.23, 2.24), and the eigenvalues of a matrix depend continuously on its entries, we know that the matrices A_u, A_c will be stable for all parameter settings sufficiently close to those special cases. Therefore, there exists a non-trivial parameter domain of local stability. It seems reasonable to conjecture that at least local stability holds for *any* set of $\mu_{\mathcal{E}}$, provided Assumption 2.4 holds. This seems a particularly natural assumption given that the graph of available routing choices has no cycles along which an instability could propagate and grow.

However, this intuition turns out to be false. We will now construct examples to demonstrate that, in general, the system can be locally unstable.

EXAMPLE 2.32 (Local instability: underload). Consider a system with 3 customer types A, B, C and 4 server types 1 through 4, connected $1 - A - 2 - B - 3 - C - 4$. Set $\beta_1 = 0.97$ and $\beta_2 = \beta_3 = \beta_4 = 0.01$. Set $\mu_{A1} = \mu_{B2} = \mu_{C3} = 1$, and $\mu_{A2} = \mu_{B3} = \mu_{C4} = 100$. (See Figure 2.6.)

For this example, we compute using Lemma 2.23,

$$A_u = \begin{pmatrix} -1.99 & -0.99 & -.99 \\ 97.02 & -2.98 & -1.98 \\ 96.03 & 96.03 & -3.97 \end{pmatrix}$$

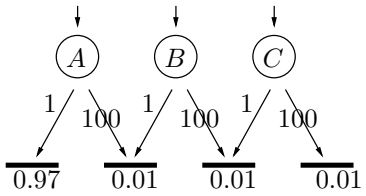

FIGURE 2.6. System with three customer types whose underload equilibrium is unstable.

with eigenvalues $\{-17.8, 4.45 \pm 23.4i\}$. Therefore by Theorem 2.21, the system with these parameters is described by an unstable ODE in the neighbourhood of its equilibrium point.

REMARK 2.33. For this, as for any other, set of parameter values $\mu_{\mathcal{E}}$, $\beta_{\mathcal{J}}$, there exist values $\lambda_{\mathcal{I}}$ which make all the activities in $\mathcal{E}$ basic. For example, we may simply start with a load-balancing allocation $\psi_{\mathcal{E}}$, and define $\lambda_{\mathcal{I}} = \sum_j \mu_{ij}\psi_{ij}$. Since the stability of A_u and A_c does not depend on the arrival rates $\lambda_{\mathcal{I}}$ (as long as the basic activity tree is unchanged), it does not matter which of the possible arrival rates we choose.

REMARK 2.34. Although we have shown that for the parameters in Example 2.32, fluid models on a neighbourhood of equilibrium are governed by an unstable ODE (and will see in Section 7 that the stochastic system is never very close to the equilibrium point), this leaves open the question of the steady-state behaviour of fluid models. In principle, as Remark 2.15 shows, it is possible to construct the unique explicit solution to the fluid model equations (with the added constraint $\lambda_{ij}(t) \geq 0$); however, as Lemma 2.35 shows, we must be dealing with a system of dimension ≥ 6 (and, it seems, with somewhat unwieldy parameters), which makes the numerical analysis somewhat involved.

On general grounds, we can conclude that, for $\rho < 1$, all fluid limits started in a compact set K will reside in some other compact set K'. This follows from the arguments in Remark 2.27, whose contents are essentially as follows: if we look at a LQFS-LB system over a sufficiently long time scale, and it starts with a large queue size, then after a finite time all server pools will be fully busy, and hence will be processing workload faster than it arrives. For the (deterministic) fluid limit, it is in principle possible to give exact bounds of the form "If the initial queue size satisfies $\|q_{\mathcal{I}}(0)\| > Q$, then after a finite time T_0 all server pools will be fully busy for at least another time T_1 such that $\|q_{\mathcal{I}}(T_0 + T_1)\| < \|q_{\mathcal{I}}(0)\|$." (The precise values of Q, T_0, and T_1 are unenlightening.) This means that there are three possible behaviours for the fluid model equations in the long run:

(1) It is possible that all fluid models eventually hit the submanifold of convergence of the linear ODE that governs the evolution of the system near the equilibrium point, and thus eventually they converge to the equilibrium point. (This seems unlikely.)
(2) The fluid model solutions may be periodic, or may converge to some periodic solution (which does not enter the region near the equilibrium point).
(3) The fluid model solutions may be chaotic. This intuitively seems like the most likely possibility, at least with generic parameter values.

This instability example is minimal in the following sense.

LEMMA 2.35. *Consider an underloaded system, $\rho < 1$.*

(1) *Let $I \geq 2$. Any customer type i that is a leaf in the basic activity tree, does not affect the local stability of the system. Namely, let us modify the system by removing type i, and then modifying (if necessary) input rates λ_k of the remaining types $k \in \mathcal{I} \setminus i$ so that the basic activity tree of the modified system is $\mathcal{E} \setminus (ij)$, where (ij) is the (only) edge in $\mathcal{E}$ adjacent to i. Then, the original system is locally stable if and only if the modified one is.*

(2) *A system with two (or one) non-leaf customer types is locally stable.*

PROOF. (1) If customer type i is a leaf, the equation for $\psi_i(t)$ is simply $\dot{\psi}_i(t) = \lambda_i - \mu_{ij}\psi_i(t)$. This means that the unit vector in the i^{th} coordinate direction is an eigenvector of A_u with the corresponding eigenvalue $-\mu_{ij} < 0$. Further, it is easy to see that: (a) the rest of the eigenvalues of A_u are those of matrix $A_u^{(-i)}$ obtained from A_u by removing the ith row and the ith column; and (b) $A_u^{(-i)}$ is exactly the "A_u-matrix" for the modified system.

(2) We can assume that there are no customer type leaves. The case $I = 1$ is trivial (and is covered by Theorem 2.30), so let $I = 2$. Throughout the proof, the pool sizes β_j are fixed. From Theorem 2.30 we know that for a certain set of service rate values (namely, $\mu_{ij} = \mu_j$, $(ij) \in \mathcal{E}$), the matrix A_u is stable. Suppose that we continuously vary the parameters μ_{ij} from those initial values to the values of interest, without ever making $\mu_{ij} = 0$. If we assume that the final matrix A_u is *not* stable, then as we change μ_{ij} the (changing) matrix A_u acquires at some point two purely imaginary eigenvalues. In that case, we must have $\text{trace}\,(A_u) = 0$. However, as seen from the form of A_u in Lemma 2.23, the diagonal entries of A_u are always negative, and therefore $\text{trace}\,(A_u) < 0$. The contradiction completes the proof.

□

This argument explains how the parameters in Example 2.32 were chosen. For 3 customer types, let the characteristic polynomial of A_u be $x^3 - c_2x^2 + c_1x - c_0$. A necessary and sufficient condition for all roots of the polynomial to have negative real parts is: $-c_2, c_1, -c_0 > 0$ and $c_2c_1 < c_0$ (see [Farkas, 2001, A1.1.1]). Using Lemma 2.23, we can evaluate the expression $c_0 - c_1c_2$ in terms of the system parameters, and look for terms which appear with a "−" sign. Setting the corresponding parameters to be large relative to the rest will produce a candidate parameter set. Appendix D.1 contains the computations.

EXAMPLE 2.36. It is possible to construct an instability example with more reasonable values of $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$, although it will have more customer types. Figure 2.7 shows the diagram. The associated 21×21 matrix A_u may be found in Appendix D.2; its largest eigenvalue has real part $\approx 0.03 > 0$.

REMARK 2.37. One of the justifications used in Remark 2.8 for the assumption that the basic activity tree $\mathcal{E}$ is known in advance was an argument of separation of time scales; the routing of customers happens over quite short time scales. One could therefore argue that the rather slow exponential growth of the instability caused by an eigenvalue with real part 0.03 is irrelevant. This intuition, however, is somewhat difficult to quantify.

EXAMPLE 2.38 (Local instability: critical load, $q > 0$). We now analyse the critically loaded system $\rho = 1$ with queues, i.e. the stability of the matrix A_c. Recall that the transformation A_c, restricted to subspace $\mathcal{N} \equiv \{y : \sum_i y_i = 0\}$, and hence the stability of A_c, does not depend on $\beta_{\mathcal{J}}$, so it suffices to specify $\mu_{\mathcal{E}}$.

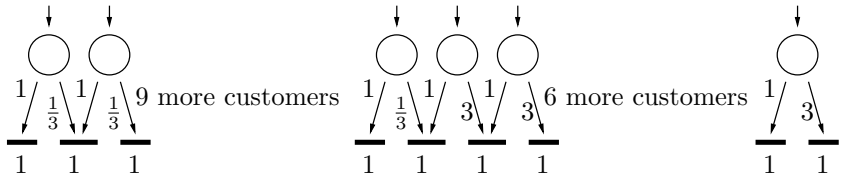

FIGURE 2.7. System with $\beta_j = 1$ and $\mu_{ij} \in \{\frac{1}{3}, 1, 3\}$ whose underload equilibrium is unstable. There are 21 customer types; $\mu_{ij} = 1$ for edges going to the left, $\mu_{ij} = \frac{1}{3}$ for the first 12 edges going to the right, $\mu_{ij} = 3$ for the last 9 edges going to the right.

Consider the network of Figure 2.8, which has 5 customer types A through E and 4 server types 1 through 4, connected $A-1-B-2-C-3-D-4-E$, with the following parameters:

$$\begin{array}{llll} \mu_{A1} = 1 & \mu_{B1} = 100 & \mu_{B2} = 1 & \mu_{C2} = 100 \\ \mu_{C3} = 1 & \mu_{D3} = 100 & \mu_{D4} = 10000 & \mu_{E4} = 100 \end{array}$$

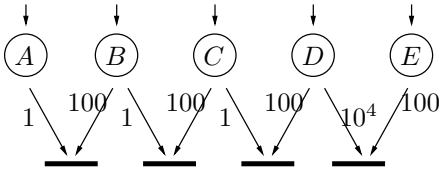

FIGURE 2.8. System with five customer types whose critical load equilibrium is unstable

The matrix A_c, computed from Lemma 2.24 will be given by

$$A_c = \frac{1}{20}\begin{pmatrix} 9389 & 9805 & 10201 & 10597 & -29003 \\ 10894 & 9290 & 9706 & 10102 & -29498 \\ 10399 & 10795 & 9191 & 9607 & -29993 \\ -40091 & -39695 & -39299 & -40903 & 119497 \\ 9409 & 9805 & 10201 & 10597 & -31003 \end{pmatrix}$$

and the eigenvalues of A_c are $\{0, -16.88, -2190.05, 2.565 \pm 23.23i\}$.

Again, the above example is in a sense minimal:

LEMMA 2.39. *Consider a critically loaded system, $\rho = 1$.*

(1) *Let $J \geq 2$. Any server type j that is a leaf in the basic activity tree does not affect the local stability of the system. Namely, let us modify the system by removing type j, and then replacing λ_i for the unique i adjacent to j by $\lambda_i - \beta_j\mu_{ij}$. Then, the original system is locally stable if and only if the modified one is.*

(2) *Consider a system labelled S. We say that a system S' is an* expansion *of system S if it is obtained from S by the following modification. We pick one server type j and one customer type i adjacent to it in $\mathcal{E}$; we "split" type j into two types j' and j''; we "connect" type i to both j' and j''; each of the remaining types $i' \in \mathcal{C}(j) \setminus i$ we connect to either j' or j'' (but not both); if $(i'j')$ (respectively $(i'j'')$) is a new edge, we set $\mu_{i'j'} = \mu_{i'j}$ (respectively $\mu_{i'j''} = \mu_{i'j}$.) Then, S is locally stable if and only if S' is.*

(3) *A system with four or fewer customer types is locally stable.*

PROOF. (1) The argument here is similar to the one used to show the independence of transformation A_c (restricted to $(I-1)$-dimensional invariant subspace)

from $\beta_{\mathcal{J}}$ in the proof of Lemma 2.24. Namely, it is easy to check that the original and the modified system share exactly the same ODE (18).

(2) Again, it is easy to see that the two systems share the same ODE (18).

(3) We can assume that there are no server-type leaves, so that the tree $\mathcal{E}$ has only customer-type leaves, of which it can have two, three, or four. We now classify these trees.

If it has four customer type leaves, then the tree has a total of four edges, hence five nodes, i.e. a single server pool, to which all the customer types are connected.

If the tree has three customer type leaves, then letting k be the number of edges from the fourth customer type, we have $k+3$ total edges, so $k+4$ nodes, of which k are server types. That is, the non-leaf customer type is connected to all of the server types. Since there are no server type leaves, we must have $k \leq 3$; since we are assuming the fourth customer type is not a leaf, we must have $k \geq 2$; thus, $k = 2$ or $k = 3$.

The last case is of two customer type leaves. Letting k, l be the number of edges coming out of the other customer types, we have a total of $k+l+2$ edges. On the other hand, since each server type has at least 2 edges coming out of it, we have at most $(k+l+2)/2$ server types, so at most $(k+l+2)/2+4$ nodes. Thus, we have $(k+l+2)+1 \leq (k+l+2)/2+4$, or $k+l+2 \leq 6$, giving $k = l = 2$ (since they must both be ≥ 2).

We summarize the possibilities in Figure 2.9. Note that the bottom-left system can be obtained by a sequence of expansions (in the sense of (2) above) from each of the top-left systems. Applying Lemma 2.39 we find that, to establish local stability for systems with four customer types, we only need to consider two systems: bottom-left and right. In both of the resulting cases, we can use

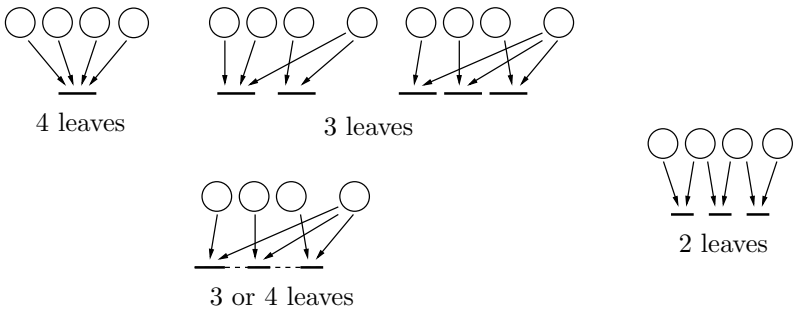

FIGURE 2.9. Possible arrangements of four customer types.

Lemma 2.24 to write out A_c and its characteristic polynomial explicitly. The characteristic polynomial will have degree 4, but one of its roots is 0, so we can reduce it to degree 3. We then symbolically verify that the stability criterion for degree 3 polynomials cited above [Farkas, 2001, A1.1.1] is satisfied. Computations can be found in Appendix D.3.

□

An argument similar to that in the above proof allows us to explain how the parameters in Example 2.38 were chosen. We seek a condition satisfied by the coefficients of a degree 4 polynomial with two imaginary roots. Letting the polynomial be $x^4 - c_1x^3 + c_2x^2 - c_3x + c_4$, and letting the roots be η_1, η_2, $\pm iz$ (where η_1 and η_2 may be real or complex conjugates, and $z \in \mathbb{R}$), we see that $c_1 = \eta_1 + \eta_2$, $c_2 = \eta_1\eta_2 + z^2$, $c_3 = (\eta_1 + \eta_2)z^2$, and $c_4 = \eta_1\eta_2 z^2$. This implies the relation $c_4c_1^2 + c_3^2 - c_1c_2c_3 = 0$, and we can find the parameters for which

this is true. (The symbolic calculation will involve rather a lot of terms, and we do not reproduce it here.)

REMARK 2.40. Whereas for polynomials of degree 3 the condition $c_2c_1 - c_0 = 0$ is both necessary and sufficient for the existence of two imaginary roots [Farkas, 2001, A1.1.1], the condition we derive here for polynomials of degree 4 is only necessary. (For example, the polynomial $(x-1)^2(x+1)^2$ has $c_1 = c_3 = 0$, so $c_4c_1^2 + c_3^2 - c_1c_2c_3 = 0$, but it has no imaginary roots.) Thus, checking the sign of the corresponding expression alone is insufficient to determine whether the system is unstable, but is a useful way of narrowing down the parameter ranges.

EXAMPLE 2.41 (Local instability: both underload and critical load ($q > 0$)). It is possible to construct a single set of parameters for which both A_u and A_c will be unstable. For the local stability of the underloaded system, the leaves of the basic activity tree corresponding to customer types are irrelevant (the corresponding occupancy on the sole available server class converges to nominal exponentially). On the other hand, for the critically loaded system, the leaves corresponding to server pools are irrelevant, since the corresponding server is fully occupied by its unique available customer type. This observation allows us to merge the above two systems into a single one which is unstable both in the underloaded and in the critically loaded case.

Consider a system with 5 customer types A through E and 5 server types 0 through 4 connected as $0 - A - 1 - B - 2 - C - 3 - D - 4 - E$, with $\mu_{A0} = 100$ and the remaining μ_{ij} as in the critically loaded case. Set $\beta_3 = 0.96$ while $\beta_0, \beta_1, \beta_2, \beta_4 = 0.01$. (See Figure 2.10.) By the above discussion, this system, which is a modification of Ex-

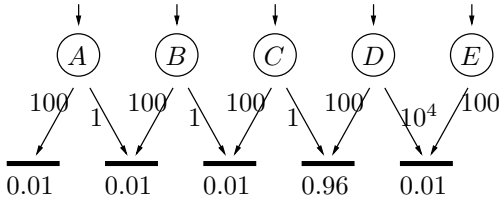

FIGURE 2.10. System with five customer types whose underload and critical load equilibrium points are both unstable

ample 2.38, must be unstable for $\rho = 1$ and positive queues, with the same eigenvalues $\{0, -16.88, -2190.05, 2.565 \pm 23.23i\}$. On the other hand, in underload, we construct the matrix A_u. We may restrict A_u to the first 4 customer types, since E is a customer type leaf and, by Lemma 2.35 (1) doesn't matter for the stability of the system.

$$A_u = \begin{pmatrix} -1.99 & -0.99 & -0.99 & -0.99 \\ 97.02 & -2.98 & -1.98 & -1.98 \\ 96.03 & 96.03 & -3.97 & -2.97 \\ -99 & -99 & -99 & -199 \end{pmatrix}$$

with eigenvalues $\{-14.6, -201.1, 3.91 \pm 18.1i\}$.

REMARK 2.42. Although for this system, both A_u and A_c are unstable, it is not obvious what the system behaviour would be like in the vicinity of the $q = 0$ equilibrium point: the trajectories governed by either matrix cross the $\rho = 1, q = 0$ boundary, and the question of stability of such hybrid systems is in general quite difficult. (For example, it is certainly possible to have two individually-unstable matrices combine to form a stable system.)

EXAMPLE 2.43 (Local instability: common eigenvector in underload and critical load). The expression (17) suggests another way to construct a system which is always locally unstable. Namely, we find a set of parameters for which A_u has a right eigenvector $(1,\dots,1)^\top$ with some non-zero real eigenvalue c, and such that $A_c = \pi A_u$ is unstable (where π is the projection along $(1,\dots,1)^\top$). Then the projection of the system state onto the manifold $\mathcal{N}$ defined in Lemma 2.24 will always evolve according to A_c, which is unstable. (See §7.3.)

Specifically, consider the system diagrammed in Figure 2.11. For sufficiently small ϵ, the matrix A_c for this system will be unstable, because the system in Figure 2.8 was unstable (i.e., had an eigenvalue with a positive real part), and the eigenvalues of a matrix depend continuously on its entries. By Lemma 2.39 (1), the addition of the server-type leaves 0 and 5 does not change critical-load stability.

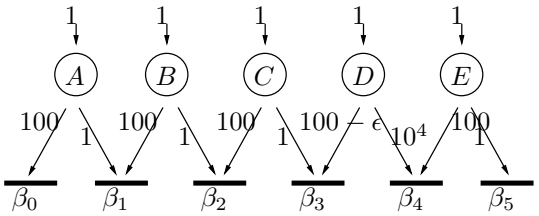

FIGURE 2.11. Modification of example in Figure 2.8, for which A_u will have $(1,\dots,1)^\top$ as an eigenvector.

We will next design the system parameters for which A_u has eigenvector $(1,\dots,1)^\top$. For this to be the case, it suffices to construct a system for which $\psi_i^* = \sum_j \psi_{ij}^*$ are all equal, and $\sum_j \mu_{ij}\psi_{ij}^* = \lambda_i = 1$ for all i. Once we find a set of suitable parameters $\psi_{ij}^* > 0$, we will set $\beta_j = \sum_i \psi_{ij}^*$ to guarantee that $\psi_{\mathcal{E}}^*$ achieve load balancing. (Recall that the linear transformation A_c, and in particular its stability properties, does not depend on $\beta_{\mathcal{J}}$ – see Lemma 2.24 (3).)

For $\delta > 0$ small, choose $\psi_{D3}^* = \frac{1-\delta}{100-\epsilon}$ and $\psi_{D4}^* = \frac{\delta}{10^4}$; then

$$\sum_j \mu_{Dj}\psi_{Dj}^* = 1, \quad \psi_D^* = \psi_{D3}^* + \psi_{D4}^* > \frac{1}{100}.$$

(This requires changing μ_{D3} from 100 to $100-\epsilon$, otherwise we could not get $\psi_{D4}^* > 0$.) Next, set $\psi_{A0}^* = \frac{1}{100} - \delta_1$ and $\psi_{A1}^* = 100\delta_1$, with $\delta_1 > 0$ small, so that $\psi_A^* = \psi_D^*$. Set $\psi_{E4}^* = \psi_{C2}^* = \psi_{B1}^* \equiv \psi_{A0}^*$ and $\psi_{E5}^* = \psi_{C3}^* = \psi_{B2}^* \equiv \psi_{A1}^*$.

We have shown the following

PROPOSITION 2.44. *There exists a set of parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$ for which the following hold:*

1. *Assumption 2.4 holds, and the unique optimal solution to the static planning problem* (1) *has $\rho = 1$.*
2. *The matrix A_c is unstable.*
3. *The matrix A_u has the vector $(1,\dots,1)^\top$ as a right eigenvector, with some real (non-zero) eigenvalue.*

In the above construction, the eigenvalue in question will be given by $\sum_j \psi_{ij}^*$.

We have shown that systems with two customer types cannot be locally unstable, for any parameter setting such that $\rho < 1$ (and Assumption 2.4 is satisfied). Here we show a form of converse to this result, namely, that for large systems there always are parameters rendering the system unstable.

LEMMA 2.45. *Let $\rho < 1$. Any shape of basic activity tree that includes a locally unstable system (i.e., with A_u having an eigenvalue with positive real part) as a subset will, with some set of parameters $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$, become locally unstable. In particular, any shape of basic activity tree that includes Example 2.32 will be locally unstable for some set of parameters $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$.*

PROOF. Let U be any system whose underload ($\rho < 1$) equilibrium is locally unstable, e.g. one of the examples given above, with the associated fixed set of parameters μ_{ij}, β_j and λ_i. Let S be a system including U as a subset, namely: the activity tree of S is a superset of that of U; the μ_{ij} and β_j in U are preserved in S; the μ_{ij} in S are fixed. Consider a sequence of systems S^ϵ in which $\beta_j = \epsilon \to 0$ for all j not in U. By Remark 2.33, for each ϵ, we can slightly perturb the arrival rates λ_i to λ_i^ϵ, such that as $\epsilon \to 0$ we have convergence

$$\lambda_i^\epsilon \to \begin{cases} \lambda_i, & i \in U \\ 0, & i \notin U \end{cases}$$

and all of the activities in S^ϵ are basic. (Keeping the value ρ from the system U, we simply prescribe the desired new occupancies ψ_{ij}^ϵ; $\beta_j \to 0$ implies $\psi_{ij}^\epsilon \to 0$ as $\epsilon \to 0$, so λ_i^ϵ will have the desired convergence properties.)

Order the ψ_i so that the customer types i in U come first. Suppose there are I customer types in U and $I+k$ customer types in S. Let A_u^ϵ be the $(I+k)\times(I+k)$ matrix associated with S^ϵ, and let A_u be the $I \times I$ matrix associated with U considered as an isolated system. Then as $\epsilon \to 0$ the top left $I \times I$ entries of A_u^ϵ converge to A_u, while the bottom left $k \times I$ entries of A_u^ϵ converge to 0. (That is, the effect of U on the stability of the rest of the system vanishes – this is due to the fact that pool size parameters β_j in U remain constant, while $\beta_j \to 0$ in the rest of the system.) Consequently, each eigenvalue of A_u is a limit of eigenvalues of A_u^ϵ. Since A_u had an eigenvalue with positive real part, for sufficiently small ϵ the matrix A_u^ϵ will have at least one eigenvalue with positive real part as well, so the system S^ϵ will be locally unstable. □

We do not have an explicit characterization of the local instability domain, either for the underloaded case or for the critically loaded one, beyond the necessity of $I \geq 3$, respectively $I \geq 4$. We informally conjecture that the phenomenon is "rare":

CONJECTURE 2.46 (Very informal). All examples of instability have somewhat unrealistic parameters (either very many customer classes, or widely differing server pool sizes or service rates).

6. LAP fluid models: convergence to equilibrium

We now switch our attention to the fluid models for the Leaf Activity Priority (LAP) algorithm, described by (7). For LAP, we only treat the case of strict underload, $\rho < 1$.

The LAP discipline is not designed with load balancing in mind; consequently, its equilibrium point is different from the load-balancing one. Instead, we recursively define the "routing rates" $\lambda_{ij} \geq 0$ as follows. For the activity $(1j)$ with the highest priority, define either $\lambda_{1j} = \lambda_1$ and $\psi_{1j}^* = \frac{\lambda_1}{\mu_{1j}}$, or $\psi_{1j}^* = \beta_j$ and $\lambda_{1j} = \beta_j \mu_{1j}$, according to whichever is smaller. Replace λ_1 by $\lambda_1 - \lambda_{1j}$ and β_j by $\beta_j - \psi_{1j}^*$, and remove the edge $(1j)$ from the tree. Proceed similarly with the remaining activities. Formally, we define the equilibrium point as follows.

DEFINITION 2.47. Assume $\rho < 1$. Set

$$\lambda_{ij} = \min\left(\lambda_i - \sum_{j':(ij')<(ij)} \lambda_{ij'}, \mu_{ij}\left(\beta_j - \sum_{i'<i} \frac{\lambda_{i'j}}{\mu_{i'j}}\right)\right).$$

Since the definition is in terms of higher-priority activities, this defines ($\lambda_{\mathcal{E}}$ uniquely. The *LAP equilibrium point* is defined to be the vector

$$(\psi^*_{\mathcal{E}}, q^*_{\mathcal{I}})$$

given by

$$\psi^*_{ij} = \frac{\lambda_{ij}}{\mu_{ij}}, \quad q^*_i = 0 \text{ for all } (ij) \in \mathcal{E},\ i \in \mathcal{I}. \tag{21}$$

(Since we're always in the underloaded case $\rho < 1$, it makes sense that all queues should be 0 at equilibrium.)

It is easy to check that, if $\rho < 1$, we will not "run out of room" halfway through the process: given the existence of some allocation $\psi_{\mathcal{E}}$ given by the solution to (1) with $\sum_i \psi_{ij} < \beta_j$ for all $j \in \mathcal{J}$, assigning the $\psi_{\mathcal{E}}$ as above instead will create enough room for all the customers. To avoid trivial complications we make the following assumption:

ASSUMPTION 2.48. *If $\psi_{\mathcal{E}}$ is a vector satisfying*

$$\psi_{ij} \geq 0,\ \forall (ij) \in \mathcal{E}, \quad \lambda_i = \sum_j \mu_{ij}\psi_{ij},\ \forall i \in \mathcal{I},$$

and

$$\sum_i \psi_{ij} \leq \beta_j\ \forall j \in \mathcal{J},$$

then $\psi_{ij} > 0$ for all $(ij) \in \mathcal{E}$. In particular, the equilibrium point satisfies this condition and, moreover, it is such that

$$\sum_i \psi^*_{ij} = \beta_j,\ \forall j < J; \quad \sum_i \psi^*_{iJ} < \beta_J.$$

The assumption means that the system needs to employ (on average) all activities in order to be able to handle the load. It holds, for example, whenever ρ is sufficiently close to 1.

REMARK 2.49. Although the LAP equilibrium point doesn't attain load balancing, the difference is negligible when system is heavily loaded (i.e. ρ is close to 1): the LAP equilibrium point is such that all queues are small and all servers are almost fully loaded, which is the best any "load balancer" could do in a heavily loaded system.

We now show that, unlike the LQFS-LB, the LAP discipline accomplishes convergence of fluid models to equilibrium for all parameter settings.

PROPOSITION 2.50. *Suppose Assumption 2.48 holds. For any $\epsilon' > 0$ and any $K > 0$ there exists a finite time $T = T(K)$ such that all fluid models whose starting state satisfies $\|\psi_{\mathcal{E}}(0), q_{\mathcal{I}}(0)\| \leq K$ have $\sum_i \psi_{ij}(t) = \beta_j,\ \forall j < J$, $q_i(t) = 0,\ \forall i \in \mathcal{I}$, and $\left|\psi_{ij}(t) - \psi^*_{ij}\right| < \epsilon'$ for all $(ij) \in \mathcal{E}$, for all $t \geq T(K)$.*

SKETCH OF PROOF. For the highest priority activity $(1j)$ there are two cases. If type 1 is a leaf, then it is easily seen that $\psi_{1j}(t) \to \psi^*_{1j}$ exponentially fast, uniformly on the initial states (bounded as in the proposition statement); this in turn implies that, after some time $T_0 = T_0(K)$, $q_1(t)$ has to be equal to 0. If pool j is a leaf, then, after some time $T'_0 = T'_0(K)$, $\psi_{1j}(t) = \psi^*_{1j} = \beta_j$. In either case, for arbitrarily small $\delta > 0$,

there exists $T_1 = T_1(\delta)$ such that $|\psi_{1j}(t) - \psi^*_{1j}| < \delta$. We can now essentially remove the highest priority activity from the tree, and proceed by induction on activity priority. (Assumption 2.48 guarantees that, for sufficiently small δ, the remaining tree will always remain connected.) $\square$

For large systems, we can use this finite time horizon result to show that sufficiently large stochastic systems will be stable, i.e. positive recurrent. (This was trivial for LQFS-LB, see Remark 2.27, but may not be true for small systems running LAP.) Moreover, in steady-state the system will sit close to equilibrium on the fluid scale.

THEOREM 2.51. *For all sufficiently large r, the LAP discipline stabilizes the network (in the sense of positive recurrence of the underlying Markov process). Moreover, the sequence of invariant distributions of $(\psi^r_{\mathcal{E}}(\cdot), q^r_{\mathcal{I}}(\cdot))$ is tight, and*

$$(\psi^r_{\mathcal{E}}, q^r_{\mathcal{I}}) \xrightarrow{w} (\psi^*_{\mathcal{E}}, q^*_{\mathcal{I}} = 0),$$

where $(\psi^*_{\mathcal{E}}, q^*_{\mathcal{I}})$ is the equilibrium point specified in Definition 2.47.

Note that the convergence in law is here convergence in probability, since the limit is a single point.

We will be using Foster-Lyapunov criteria (see e.g. [Bramson, 2006, Proposition 4.5], and references therein) to conclude stability and tightness of the associated measures. In order to do that, we need to establish some results on the behaviour of all sufficiently *large* systems over a finite time horizon.

LEMMA 2.52. *There exists $T_1 > 0$ such that for any $T_2 > T_1$ there exists sufficiently large $C = C(T_2)$ for which the following holds. For any $\epsilon > 0$,*

$$\mathbb{P}\left\{\left|\sum_{(ij)\in\mathcal{E}} \nu_i(d^r_{ij}(T_2) - d^r_{ij}(T_1)) - (T_2 - T_1)\sum_{j\in\mathcal{J}} \alpha_j\right| \geq \epsilon\right\} \to 0,$$

as $r \to \infty$, uniformly on initial states with $\max_{i\in\mathcal{I}} q^r_i(0) \geq C$. Here, ν_i and α_j are the workload (of a job of type i) and the rate of processing workload (by the server pool j) respectively, defined by (2).

In plain words, we are asserting that, if the initial state of the system is large enough, then after one finite time T_1 (uniform) and until another finite time T_2 (which grows to infinity with the starting state) the system will be processing workload at the maximal possible rate $\sum_j \alpha_j(t)$.

PROOF. The proof uses fluid models with infinite initial states. We cannot appeal directly to the properties of "standard" fluid models defined earlier, because we require convergence that is uniform in all large initial states. Instead, we consider the following version of the fluid limit result. Consider a sequence of initial states $(\psi^r_{\mathcal{E}}(0), q^r_{\mathcal{I}}(0))$ with $\|(\psi^r_{\mathcal{E}}(0), q^r_{\mathcal{I}}(0))\| = C'(r) \to \infty$ as $r \to \infty$. If we regard $q^r_{\mathcal{I}}(0) \in \overline{\mathbb{R}^I} \equiv \mathbb{R}^I \cup \{\infty\}$, any such sequence has a convergent subsequence; we will restrict our attention to such a subsequence. The condition $\|(\psi^r_{\mathcal{E}}(0), q^r_{\mathcal{I}}(0))\| \to \infty$ as $r \to \infty$ means that the limit $(\psi_{\mathcal{I}}(0), q_{\mathcal{I}}(0))$ will have $q_i(0) = \infty$ for at least one customer class i. Partition the customer classes as $\mathcal{I} = \mathcal{I}^\infty \cup \mathcal{I}^0$, where $q_i(0) = \infty$ for $i \in \mathcal{I}^\infty$, and $q_i(0) < \infty$ for $i \in \mathcal{I}^0$. Now, we can prove a fluid limit result, analogous to Proposition 2.50. Namely, with probability 1, any subsequence of fluid-scaled trajectories has a further subsequence which converges u.o.c. to a fluid model satisfying conditions (7), except that for all $i \in \mathcal{I}^\infty$ the queue length $q_i(t) = \infty, \forall t \geq 0$. These infinite-initial-state fluid models are such that, uniformly on all of them, starting at some finite time T_1', all server pools are fully occupied. Indeed, the same analysis as in Proposition 2.50 gives $\sum_i \psi_{ij}(t) = \beta_j$ for all $j < J$ after a finite

time; but the existence of infinite queues together with Assumption 2.48 guarantees that after some further finite time T_1' we will always have $q_I > 0$, so $\sum_i \psi_{iJ}(t) = \beta_J$ as well.

We choose $T_1 = 2T_1'$. Consider any $T_2 > T_1$. If the lemma were false, then for some fixed $\epsilon' > 0$ we could find a sequence of systems with $\|(\psi_{\mathcal{E}}^r(0), q_{\mathcal{I}}^r(0))\| = C'(r) \to \infty$, such that

$$\limsup_r \mathbb{P}\left\{\left|\sum_{(ij)} \nu_i(d_{ij}^r(T_2) - d_{ij}^r(T_1)) - \sum_j \alpha_j(t)(T_2 - T_1)\right| \geq \epsilon'\right\} > 0.$$

This, however, is impossible because, from the fluid limit result, we must have w.p.1.

$$\sup_{t\in[T_1,T_2]} \max_j \left|\sum_i \psi_{ij}^r(t) - \beta_j\right| \to 0,$$

and then

$$\left|\sum_{(ij)} \nu_i(d_{ij}^r(T_2) - d_{ij}^r(T_1)) - \sum_j \alpha_j(t)(T_2 - T_1)\right| \to 0. \qquad \square$$

Proof of Theorem 2.51. Recall that $\nu_i > 0$ is the workload associated with a single request of type i; i.e., the optimal dual variable associated with (1c) for type i (see (2)). We consider the *total workload*

$$W^r(t) = \sum_i \nu_i x_i^r(t).$$

We will argue that the quantity

$$\mathcal{L}^r(t) = (W^r(t))^2$$

will serve as a Lyapunov function for the rth system. Namely, the following property holds: there exist positive constants K, T, C_1, C_2, C_3 such that, for all sufficiently large r,

$$\text{if } \mathcal{L}^r(t) > K \text{ then } \mathbb{E}[\mathcal{L}^r(t+T) - \mathcal{L}^r(t)|\mathcal{L}^r(t)] < -C_1 W^r(t) + C_2 \tag{22}$$

and

$$\text{if } \mathcal{L}^r(t) \leq K \text{ then } \mathbb{E}[\mathcal{L}^r(t+T) - \mathcal{L}^r(t)|\mathcal{L}^r(t)] < C_3. \tag{23}$$

Once we show (22)–(23), a standard application of the Foster-Lyapunov criteria [Bramson, 2006, Proposition 4.5] shows that for all sufficiently large r the system Markov process is positive recurrent, and moreover, the stationary distributions are such that $\mathbb{E}W^r = \sum_i \nu_i \mathbb{E}x_i^r$ remains uniformly (in r) bounded. This implies that the sequence $(\psi_{\mathcal{E}}^r, q_{\mathcal{I}}^r)$ is tight; hence, any subsequence has a further, convergent, sub-sub-sequence. Proposition 2.50 then implies that any convergent subsequence of invariant measures must weakly converge to the point mass at equilibrium, which concludes the proof.

It remains to show (22)–(23). First, it is easy to see that, $\forall T > 0$,

$$\mathbb{E}[W^r(t+T) - W^r(t)]^2 \text{ are uniformly bounded across all } r \text{ and } t. \tag{24}$$

This guarantees (23) for any fixed K. To prove (22), we fix $T_1 > 0$ as in Lemma 2.52, and then choose a large fixed $T > T_1$. Note that

$$(\min_{i\in\mathcal{I}} \nu_i)(\max_{i\in\mathcal{I}} q_i^r(t)) \leq W^r(t) \leq (\max_{i\in\mathcal{I}} \nu_i)(I \max_{i\in\mathcal{I}} q_i^r(t) + \sum_j \beta_j);$$

therefore, we may replace $\max_{i\in\mathcal{I}} q_i^r(0) \to \infty$ by $W^r(0) \to \infty$ in the conditions of Lemma 2.52. If we fix a sufficiently small $\epsilon' > 0$ and apply Lemma 2.52, we obtain

the following fact:
For a sufficiently large fixed $K > 0$, uniformly on all $\mathcal{L}^r(0) > K$ and all large r,

$$\mathbb{P}\left\{W^r(T) - W^r(0) \le 2\sum_i \lambda_i \nu_i T_1 - \frac{1}{2}(1-\rho)(T-T_1)\sum_j \alpha_j\right\} \ge 1 - \epsilon'. \tag{25}$$

Here, the term $2\sum_i \lambda_i \nu_i T_1$ is a crude upper bound on $W^r(T_1) - W^r(0)$, which holds with high probability for large r. The term $-\frac{1}{2}(1-\rho)(T-T_1)\sum_j \alpha_j$ is an upper bound on $W^r(T) - W^r(T_1)$, also holding with high probability, which follows from Lemma 2.52 and relation (3)).

When T is large enough, the right-hand side of the first inequality in (25) is negative. This, along with (24), implies (22). □

7. LQFS-LB steady-state on the diffusion scale

We will now analyse the behaviour of the LQFS-LB on the diffusion scale; that is, we will consider deviations of the system state from its equilibrium point (Definition 2.19), scaled down by $\sqrt{r}$. First we show that, over a finite time horizon, the behaviour of the system can be described by a diffusion process satisfying a certain stochastic differential equation. We will then consider the steady state behaviour on the same scale. We will show the following three results:

- If $\rho < 1$ and fluid models are locally unstable, then the steady state of the system does *not* live on the diffusion scale; that is, the invariant measure of a ball of size $K\sqrt{r}$ around the equilibrium point converges to 0 as $r \to \infty$, for any K. (Theorem 2.58.)
- The model with parameters satisfying Proposition 2.44 will not display "Halfin-Whitt-like" behaviour. (A summary of [Halfin and Whitt, 1981] can be found in Appendix C.) That is, if $\rho^r \to 1$ with $1 - \rho^r = O(\sqrt{r})$, the probability of an arriving call having to wait converges to 1. (Theorem 2.60.)
- If, however, the service rate depends only on the server type ($\mu_{ij} = \mu_j$ for all i), then both of the above are reversed. When $\rho < 1$, the steady-state deviations of such a model from its equilibrium point, scaled down by $\sqrt{r}$, are tight; when $\rho^r \to 1$ with $1 - \rho^r = O(\sqrt{r})$, the same tightness holds, and implies that the probability of an arriving customer having to wait has a limiting value strictly between 0 and 1. (Theorem 2.62.)

Remark 2.53. These three possibilities are not exhaustive: for example, in the case $\mu_{ij} = \mu_i$ we have shown local stability, but have not shown either global stability or lack of it. Theorem 2.62 applies whenever we have both global and local stability. Conversely, lack of global stability suggests, but by no means proves, that the invariant measure does not live near equilibrium; but we have only been able to show this when local stability fails.

We begin by defining the *diffusion scaling*.

Definition 2.54. For all state variables Γ, we let

$$\hat{\Gamma}^r(t) \equiv \frac{\Gamma^r(t) - r\gamma^*}{\sqrt{r}}.$$

Specifically, we will be interested in the quantities

$$\hat{\Psi}^r_{ij}(t) \equiv \frac{\Psi^r_{ij}(t) - r\psi_{ij}}{\sqrt{r}} \tag{26}$$

and the derived quantities

$$\hat{\Psi}_i^r(t) \equiv \sum_j \hat{\Psi}_{ij}^r(t), \quad \hat{\Psi}_j^r(t) \equiv \sum_i \hat{\Psi}_{ij}^r(t) = \frac{\Psi_j^r(t) - \rho r \beta_j}{\sqrt{r}}.$$

We will be interested in the behaviour of the system under this scaling in two settings: in underload and in the Halfin-Whitt regime. In underload, our assumptions on the set-up of the system are as before: namely, the rth system has arrival rates $\lambda_i^r \equiv \lambda_i r$, server pool sizes $\beta_j^r \equiv \beta_j r$, and service rates μ_{ij}, where the parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, and $\mu_{\mathcal{E}}$ satisfy Assumption 2.4. We now define the Halfin-Whitt regime for the multi-class case as follows:

DEFINITION 2.55. The *Halfin-Whitt asymptotic regime* is a family of systems, indexed by r, with the following properties. We consider a set of parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, and $\mu_{\mathcal{E}}$ such that the unique optimal solution to the static planning problem (1) has $\rho = 1$, and Assumption 2.4 is satisfied. In the rth system, $\beta_j^r \equiv r\beta_j$ (same as throughout the chapter). However, the input rates are

$$\lambda_i^r \equiv r\lambda_i + \sqrt{r} l_i,$$

for some set of real numbers $l_{\mathcal{I}}$ such that $\sum l_i \nu_i = -C < 0$. (Here, $\nu_{\mathcal{I}}$ are workloads defined by (2).)

Denote by $\rho^r, \{\lambda_{ij}^r\}$ the optimal solution of SPP (1) with β_j and λ_i replaced by β_j^r and λ_i^r respectively. (Under Assumption 2.4, this solution is unique for all large r.) Because ρ^r can equivalently be defined through workloads as in (3), and the workloads will be the same for all large r, we have $\rho^r = 1 + (\sum l_i \nu_i)/\sqrt{r} = 1 - C/\sqrt{r}$. This in turn implies that, for all large r, the Markov process describing the model is positive Harris recurrent, with a unique invariant distribution; so it makes sense to speak of steady-state variables.

We use the notation of (26) in the Halfin-Whitt regime, with the convention $q_i^* = 0$ and $z_j^* = 0$. Recall that $Z_j^r(t) \leq 0$ measures the idleness of server pool j, and is given by $Z_j^r(t) \equiv \Psi_j^r(t) - r\beta_j$. We note that $\hat{Z}_j^r(t)$ is measuring the deviation of pool-j occupancy from full occupancy $r\beta_j$, not from its equilibrium value in the rth system, $\rho^r r\beta_j$. Thus, we have queueing if $\hat{Q}_i^r > 0$, and we have idleness if $\hat{Z}_j^r < 0$.

Recall the load-balancing mapping $M : \mathbb{R}^I \to \mathbb{R}^{I+J-1}$ (9), which sent $\psi_{\mathcal{I}}$ to the load-balancing allocation $\psi_{\mathcal{E}}$. Let M' be its left inverse, namely,

$$M' z_{\mathcal{E}} = \left(\sum_j z_{ij} \right)_{i \in \mathcal{I}}.$$

We can rewrite the manifold $\mathcal{M}$ defined in (10) as

$$\mathcal{M} = \{ z \in \mathbb{R}^{I+J-1} : z = MM'z \}.$$

The state space collapse results of Theorems 2.21, 2.22 suggest that the queueing network should "live on" $\mathcal{M}$; we will see that this is true. Specifically, we show that if the network starts close to the equilibrium point on the diffusion scale, then under the diffusion scaling it will jump to $\mathcal{M}$ instantaneously, and then over a finite time it will evolve within $\mathcal{M}$.

7.1. Finite time horizon diffusion process approximation. We will require an approximation of the behaviour of the network under the diffusion scaling over a finite time horizon. Derivation of such behaviour is nearly standard; and, in any case, was done in Gurvich and Whitt [2009] in some generality. (Our load-balancing algorithm belongs to the family of algorithms that they consider.) The exposition below follows Gurvich and Whitt [2009] and Dai and Tezcan [2011], omitting some of the more technical details.

The term "finite time horizon" means that we will be concerned with uniform convergence of processes on compact sets. That is, in this section we fix a time interval $[0,T]$ and look at the behaviour of the rth system on it, rather than examining the steady-state behaviour. We will return to studying the steady-state behaviour in §7.2–7.4.

THEOREM 2.56 (Essentially a corollary of [Gurvich and Whitt, 2009, Theorem 3.1 and Theorem 4.4]). *Let $\rho < 1$. Assume that as $r \to \infty$,*

$$\hat{\Psi}^r_{\mathcal{E}}(0) \to \hat{\Psi}_{\mathcal{E}} \tag{27}$$

where $\hat{\Psi}_{\mathcal{E}}$ is deterministic and finite. (Consequently, $\hat{\Psi}^r_{\mathcal{I}}(0) \to \hat{\Psi}_{\mathcal{I}}(0) \equiv M'\hat{\Psi}_{\mathcal{E}}$.) Then,

$$\hat{\Psi}^r_{\mathcal{I}}(\cdot) \implies \hat{\Psi}_{\mathcal{I}}(\cdot) \text{ in } D^I[0,\infty), \tag{28}$$

and for any fixed $\eta > 0$,

$$\hat{\Psi}^r_{\mathcal{E}}(\cdot) \implies M\hat{\Psi}_{\mathcal{I}}(\cdot) \text{ in } D^{I+J-1}[\eta,\infty), \tag{29}$$

where $\hat{\Psi}_{\mathcal{I}}(\cdot)$ is the unique (possibly weak) solution of the stochastic differential equation

$$\hat{\Psi}_{\mathcal{I}}(t) = \hat{\Psi}_{\mathcal{I}}(0) + \int_0^t A_u \hat{\Psi}_{\mathcal{I}}(s)ds + \sqrt{2\lambda_i}B_i(t), \tag{30}$$

the matrix A_u is defined by (12)*, and the processes $B_i(\cdot)$ are independent standard Brownian motions.*

SKETCH OF PROOF. We will not justify why limiting processes can be defined (details can be found in Gurvich and Whitt [2009], or in [Dai and Tezcan, 2011, Theorem 4.2]). We will, however, justify why any subsequential limit must satisfy the SDE (30).

Fix an interval $[0,T]$. The finiteness of the limit $\hat{\Psi}_{\mathcal{E}}(0)$ in (27) means that under the *fluid* scaling, the initial state converges to the fluid equilibrium point. Applying Theorem 2.21, we conclude:

$$\text{as } r \to \infty,\ \Psi^r_{\mathcal{E}}(t) = r\psi^*_{\mathcal{E}} + o(r) \text{ for all } t \in [0,T].$$

In particular, since we are in underload $\rho < 1$,

$$\mathbb{P}(Q^r_i(t) > 0 \text{ for some } i \in \mathcal{I}, t \in [0,T]) \to 0$$

as $r \to \infty$. We may therefore work on the event that there is never any queueing in the system, a significant simplification relative to Gurvich and Whitt [2009]. (We also have no abandonment.)

Assuming there is no queueing on the time interval $[0,T]$, we have $\Psi^r_i(t) = X^r_i(t)$, and we can write

$$\hat{\Psi}^r_i(t) = \hat{\Psi}^r_i(0) - \sum_j \mu_{ij} \int_0^t \hat{\Psi}^r_{ij}(s)ds + \frac{1}{\sqrt{r}}\left(\Pi^{(a)}(\lambda_i rt) - \lambda_i rt\right) - \sum_j \frac{1}{\sqrt{r}}\left(\Pi^{(s)}_{ij}\left(\mu_{ij}\int_0^t \Psi^r_{ij}(s)ds\right) - \mu_{ij}\int_0^t \Psi^r_{ij}(s)ds\right).$$

The Brownian term $\sqrt{2\lambda_i}B_i(t)$ now follows from the functional central limit theorem for Poisson processes, because by Theorem 2.21 we know that on the fluid scale the trajectory is sitting at equilibrium: thus, $\sum_j \mu_{ij}\Psi^r_{ij}(s) = \lambda_i r + o(r)$.

To conclude the sketch of proof that the limiting process satisfies the linear SDE (30), we need to demonstrate that $\hat{\Psi}^r_{\mathcal{E}}(t) \approx M\hat{\Psi}^r_{\mathcal{I}}(t)$ for all $t \in [\eta,T]$, where we can choose $\eta \to 0$

as $r \to \infty$: that is, we need to establish the state space collapse of Theorem 2.21, but on the diffusion scale[10]. This is accomplished by considering the *hydrodynamic scaling*,[11] ,

$$\overline{\gamma}^{r,m}(t) = \frac{1}{\sqrt{r}} \left(\Gamma^r \Big(\frac{t}{\sqrt{r}} + \frac{m}{\sqrt{r}} \Big) - r\gamma^* \right).$$

(This is the same scaling as in §8.2, with $h(r) \equiv r^{1/2}$.) We will be considering the limits under this scaling, for $0 \le m < \sqrt{r}T$. Since this is a version of a fluid scaling, similarly to Theorem 2.13, as $r \to \infty$, any sequence of hydrodynamically-scaled processes (for a fixed m) has a subsequence which converges uniformly to a Lipschitz limit. Note, however, that in this limit quantities such as $\overline{\psi}^m_{ij}(t)$ measure the *deviations* of the occupancy process from the equilibrium, and as such can be negative.

The hydrodynamic model equations (for $\rho < 1$ and no queueing) are:

$$\overline{\psi}^m_i(t) = \overline{\psi}^m_i(0), \ \forall i \in \mathcal{I} \tag{31a}$$

$$\overline{\psi}^m_i(t) = \sum_j \overline{\psi}^m_{ij}(t), \ \forall i \in \mathcal{I} \tag{31b}$$

$$\overline{\rho}^m_j(t) = \left(\sum_i \overline{\psi}^m_{ij}(t) \right) / \beta_j, \ \forall i \in \mathcal{I} \tag{31c}$$

For any set of customer types $\mathcal{I}_* \subseteq \mathcal{I}$, and any set of server types $\mathcal{J}_* \subseteq \mathcal{J}$ such that $\overline{\rho}^m_j(t) < \overline{\rho}^m_{j'}(t)$ whenever $j \in \mathcal{J}_*$, $j' \notin \mathcal{J}_*$, and $\mathcal{C}(j) \cap \mathcal{C}(j') \cap \mathcal{I}_* \neq \emptyset$,

$$\sum_{j \in \mathcal{J}_*} \sum_{i \in \mathcal{C}(j) \cap \mathcal{I}_*} \frac{d}{dt} \overline{\psi}^m_{ij}(t) = \sum_{i \in \cup_{j \in \mathcal{J}_*} \mathcal{C}(j) \cap \mathcal{I}_*} \lambda_i - \sum_{j \in \mathcal{J}_*} \sum_{i \in \mathcal{C}(j) \cap \mathcal{I}_*} \mu_{ij} \psi^*_{ij} \tag{31d}$$

Equation (31a), which corresponds to the ordinary fluid model equation (5c) but has no explicit mention of the arrival or departure processes, arises as follows. Under our rescaling, the arrival process is simply linear, of rate $\lambda_i t$. On the hydrodynamic scale, the approximation $\Psi^r_{ij} \approx r\psi^*_{ij}$ implies that the service rate is (in the limit) precisely equal to the nominal value. Consequently, in the hydrodynamic limit the number of customers of type i in the system does not change. (This also accounts for the appearance of ψ^*_{ij} in (31d).)

From these equations it follows readily that whenever $\min_j \overline{\rho}^m_j(t) < \max_j \overline{\rho}^m_j(t)$, the difference between the largest and the smallest loads is decreasing at a rate bounded below by a constant. Consequently, after a finite hydrodynamic time (corresponding to some time of order $r^{-1/2}$ under the diffusion scaling), we will arrive at $\mathcal{M}$; and since hydrodynamic models which start in $\mathcal{M}$ cannot leave it, we will remain on $\mathcal{M}$ for the remainder of the time interval $[0, T]$. Thus, for large r, the rth diffusion-scaled system is very close to $\mathcal{M}$ on the time interval $[T_1 r^{-1/2}, T]$ for some fixed, finite T_1; and therefore, the limiting process will stay on $\mathcal{M}$ during $[\eta, T]$ for any $\eta > 0$.

This concludes the sketch of proof that the diffusion-scaled process converges to a continuous process satisfying the SDE (30). Properties of such SDEs (in particular, uniqueness of solutions) can be found in [Karatzas and Shreve, 1996, Chapter 5]. □

[10] This is sufficient only if we assume the existence of the limiting process. To prove the existence, we require the processes involved to be stochastically bounded; details can be found in [Gurvich and Whitt, 2009].

[11] The term "hydrodynamic" is not used in any technical sense; this is simply another version of "fluid-like" scaling, in which the system converges to a nearly-deterministic, rather than stochastic, process. This scaling regime has features which distinguish it both form the ordinary fluid and the local fluid limits, meriting a separate name.

The meaning of Theorem 2.56 is simple: the diffusion limit of the process $\hat{\Psi}^r_{\mathcal{I}}(\cdot)$ is such that, at initial time 0, it "instantly jumps" to the state $\hat{\Psi}_{\mathcal{E}}(0) \equiv MM'\hat{\Psi}_{\mathcal{E}}$ on the manifold $\mathcal{M}$ (where $\hat{\Psi}_{\mathcal{E}}(0) = \lim_{r\to\infty} \hat{\Psi}^r_{\mathcal{E}}(0) \equiv \hat{\Psi}_{\mathcal{E}}$ holds only if $\hat{\Psi}_{\mathcal{E}} \in \mathcal{M}$); after this initial jump, the process stays on $\mathcal{M}$ and evolves according to the stochastic differential equation (30). Our proofs of diffusion-scaled instability in §7.2 will rely on the analysis of this SDE.

Similar in spirit, but somewhat more involved in the details of proof is the following result, which holds in the Halfin-Whitt regime. Let π be the orthogonal projection along $(1,\dots,1)^\top$, and define the map F as follows: for $y \in \mathbb{R}^I$, set

$$F(y) = \begin{cases} \pi(y), & \sum_i y_i > 0 \\ y, & \sum_i y_i \le 0 \end{cases}. \tag{32}$$

THEOREM 2.57 (Corollary of [Gurvich and Whitt, 2009, Theorem 3.1 and Theorem 4.4]). *For a family of systems in the Halfin-Whitt regime (Definition 2.55), assume that*

$$\hat{X}^r_{\mathcal{I}}(0) \to \hat{X}_{\mathcal{I}}(0), \qquad \hat{\Psi}^r_{\mathcal{E}}(0) \to \hat{\Psi}_{\mathcal{E}} \tag{33}$$

where $\hat{X}_{\mathcal{I}}(0)$ and $\hat{\Psi}_{\mathcal{E}}$ are deterministic and finite. Then,

$$\hat{X}^r_{\mathcal{I}}(\cdot) \implies \hat{X}_{\mathcal{I}}(\cdot) \text{ in } D^I[0,\infty), \tag{34}$$

and for any fixed $\eta > 0$,

$$\hat{\Psi}^r_{\mathcal{E}}(\cdot) \implies MF(\hat{X}_{\mathcal{I}}(\cdot)) \text{ in } D^{I+J-1}[\eta,\infty), \tag{35}$$

where $\hat{X}_{\mathcal{I}}(\cdot)$ is the unique (possibly weak) solution of the stochastic differential equation

$$\hat{X}_{\mathcal{I}}(t) = \hat{X}_{\mathcal{I}}(0) + \int_0^t A_u F(\hat{X}_{\mathcal{I}}(s))ds + (\sqrt{2\lambda_i}B_i(t))_{i\in\mathcal{I}}, \tag{36}$$

and the processes $B_i(\cdot)$ are independent standard Brownian motions.

We refer to Gurvich and Whitt [2009] for the details of proof.

7.2. Steady state behaviour of locally unstable systems: evanescence of invariant measures in underload. In this section we show that if the matrix A_u has eigenvalues with positive real part, then the stationary distribution of the (diffusion scaled) process $\hat{\Psi}^r_{\mathcal{E}}(\cdot)$ escapes to infinity as $r \to \infty$. We will rely on the results of Theorem 2.56.

THEOREM 2.58. *Consider a set of parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$ such that Assumption 2.4 holds, and $\rho < 1$. Consider a sequence of systems with arrival rates $\lambda^r_{\mathcal{I}} \equiv r\lambda_{\mathcal{I}}$, server pool sizes $\beta^r_{\mathcal{J}} \equiv r\beta_{\mathcal{J}}$, and unscaled service rates $\mu_{\mathcal{E}}$. Denote by $\mathfrak{M}^r$ the stationary distribution of the process $\hat{\Psi}^r_{\mathcal{E}}(\cdot)$, a probability measure on $\mathbb{R}^{I+J-1}$. Let $b_K = \{z : |z| \le K\} \subset \mathbb{R}^{I+J-1}$ be the ball of radius K in $\mathbb{R}^{I+J-1}$.*

Suppose the matrix A_u defined in Lemma 2.23 has at least one eigenvalue with positive real part, and no purely imaginary eigenvalues[12]. Then for any K, $\mathfrak{M}^r(b_K) \to 0$ as $r \to \infty$.

Before we proceed with the proof, let us introduce more notation and one auxiliary result. Recall that, in fluid models, after a finite time the difference $\psi_{\mathcal{E}}(t) - \psi^*_{\mathcal{E}}$ lives on the manifold $\mathcal{M}$ defined by (10), and satisfies (in the vicinity of the equilibrium point) the linear ordinary differential equation

$$\dot{z} = (MA_uM')z, \quad z \in \mathcal{M}. \tag{37}$$

[12]The requirement of "no purely imaginary eigenvalues" is made for convenience of differentiating between strict convergence and strict divergence. It holds for generic values of $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$: that is, any set of values $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$ has an arbitrarily small perturbation $\tilde{\beta}_{\mathcal{J}}$, $\tilde{\mu}_{\mathcal{E}}$ with for which the corresponding matrix $\tilde{A}_u$ has no purely imaginary eigenvalues.

Let $\mathcal{C} \subset \mathcal{M}$ denote the submanifold of convergence of this ODE; that is, $\mathcal{C} = \{z : z(t) \to 0 \text{ as } t \to \infty\}$. We can equivalently define $\mathcal{C} = M\mathcal{C}_{\mathcal{I}}$, where $\mathcal{C}_{\mathcal{I}}$ is the submanifold of convergence of the linear ODE

$$\dot{y} = A_u y, \quad y \in \mathbb{R}^I;$$

this ODE describes the evolution of the fluid model quantity $\psi_{\mathcal{I}}(t) - \psi^*_{\mathcal{I}}$ near the equilibrium point.

Given assumptions of the theorem on A_u, the solutions to (37) converge to 0 exponentially fast if $z(0) \in \mathcal{C}$, which is a submanifold of positive codimension. Solutions started from points $z(0) \in \mathcal{M} \setminus \mathcal{C}$ diverge to infinity exponentially quickly[13].

We will write

$$b_K(\delta_1, \delta_2) \equiv b_K \cap \{z : d(z, \mathcal{M}) \leq \delta_1, d(z, \mathcal{C}) \geq \delta_2\},$$

where $d(\cdot, \cdot)$ is Euclidean distance.

LEMMA 2.59. *Solutions of the stochastic differential equation* (30) *have the following properties.*

(1) *For any $T > 0$ and any $\Psi_{\mathcal{I}}(0)$,*

$$\mathbb{P}\{M\hat{\Psi}_{\mathcal{I}}(T) \in \mathcal{M} \setminus \mathcal{C}\} = 1;$$

(2) *For any $K > 0$, $\delta_2 > 0$ and $\epsilon > 0$, there exist sufficiently large T_K and $K' > K$, such that, uniformly on $M\hat{\Psi}_{\mathcal{I}}(0) \in b_K(0, \delta_2)$,*

$$\mathbb{P}\{M\hat{\Psi}_{\mathcal{I}}(T_K) \in b_{K'} \setminus b_{2K}\} \geq 1 - \epsilon.$$

PROOF. Statement (1) follows from the fact that, regardless of the (deterministic) initial state $\Psi_{\mathcal{I}}(0)$, the solution to SDE (30) is such that the distribution of $\Psi_{\mathcal{I}}(T)$ is Gaussian with non-singular covariance matrix. (See [Karatzas and Shreve, 1996, Section 5.6]. In our case the matrix of diffusion coefficients is diagonal with entries $\sqrt{2\lambda_i}$.) Consequently, the probability that the state belongs to the submanifold $\mathcal{C}$ of positive codimension is 0.

Statement (2) follows from the fact [Karatzas and Shreve, 1996, Section 5.6] that the expectation $m(t) = \mathbb{E}\hat{\Psi}_{\mathcal{I}}(t)$ evolves according to the ODE

$$\dot{m}(t) = A_u m(t).$$

Since $d(M\hat{\Psi}_{\mathcal{E}}(0), \mathcal{C}) \geq \delta_2$ (and thus $\hat{\Psi}_{\mathcal{I}}(0)$ is also separated by a positive distance from $\mathcal{C}_{\mathcal{I}}$), we have

$$|m(t) - m(0)| \geq a_1 \exp(at)$$

for some fixed $a_1, a > 0$ and all large t. It is easy to see that if the mean of a Gaussian distribution goes to infinity, then (regardless of how the covariance matrix evolves) the measure of any bounded set goes to zero, so choosing T_K sufficiently large, we will have arbitrarily high probability of leaving the set b_{2K}. On the other hand, both $m(t)$ and the covariance matrix remain bounded for all $t \in [0, T_K]$, for any T_K; so K' can always be chosen sufficiently large so that $\mathbb{P}\{M\hat{\Psi}_{\mathcal{I}}(T_K) \in b_{K'}\}$ is arbitrarily close to 1. □

We are now in position to prove Theorem 2.58.

[13]If A_u has purely imaginary eigenvalues, there is a further submanifold $\tilde{\mathcal{C}}$ on which solutions orbit around the equilibrium point. The important thing for us is that the set of initial conditions for which solutions do *not* diverge to infinity is a submanifold of $\mathcal{M}$ with positive codimension, hence of measure 0.

PROOF OF THEOREM 2.58. We will treat $\mathfrak{M}^r$ as measures on the one-point compactification $\overline{\mathbb{R}^n} = \mathbb{R}^n \cup \{*\}$ of the space $\mathbb{R}^n$, where $n = I+J-1$. In this space, any subsequence of $\{\mathfrak{M}^r\}$ has a further subsequence, along which $\mathfrak{M}^r \xrightarrow{w} \mathfrak{M}$ for some probability measure $\mathfrak{M}$ on $\overline{\mathbb{R}^n}$. We will show that the entire measure $\mathfrak{M}$ is concentrated on the infinity point $*$, i.e. $\mathfrak{M}(\mathbb{R}^n) = 0$. Suppose not, i.e. $\mathfrak{M}(\mathbb{R}^n) > 0$. The proof proceeds in two steps.

Step 1. We prove that $\mathfrak{M}(\mathbb{R}^n) = \mathfrak{M}(\mathcal{M} \setminus \mathcal{C})$. Indeed, let us choose any $\epsilon > 0$, and K large enough so that $\mathfrak{M}(b_{K/2}) > (1-\epsilon)\mathfrak{M}(\mathbb{R}^n)$. Then, for all large r, $\mathfrak{M}^r(b_K) > (1-\epsilon)\mathfrak{M}(\mathbb{R}^n)$. Choose $\delta_1 > 0$ and $T > 0$ arbitrary. By Lemma 2.59, we can choose a sufficiently small $\delta_2 > 0$ and a sufficiently large K' such that, uniformly on the initial states $\hat{\Psi}^r_{\mathcal{E}}(0) \in b_K$,

$$\liminf_{r\to\infty} \mathbb{P}\{\hat{\Psi}^r_{\mathcal{E}}(T) \in b_{K'}(\delta_1, \delta_2)\} > 1 - \epsilon.$$

This implies that for all large r,

$$\mathfrak{M}^r(b_{K'}(\delta_1, \delta_2)) > (1-\epsilon)^2 \mathfrak{M}(\mathbb{R}^n),$$

and then $\mathfrak{M}(b_{K'}(\delta_1, \delta_2)) \geq (1-\epsilon)^2\mathfrak{M}(\mathbb{R}^n)$. Since ϵ and δ_1 were arbitrary, we conclude that $\mathfrak{M}(\mathbb{R}^n) \leq \mathfrak{M}(\mathcal{M} \setminus \mathcal{C})$, and then, obviously, the equality must hold.

Step 2. We show that, for arbitrarily large $K > 0$, $\mathfrak{M}(\mathbb{R}^n \setminus b_K) = \mathfrak{M}(\mathbb{R}^n)$. (This is, of course, impossible when $\mathfrak{M}(\mathbb{R}^n) > 0$, and thus we obtain a contradiction.) It suffices to show that for any $\epsilon > 0$, we can choose a sufficiently large K, such that $\mathfrak{M}(\mathbb{R}^n \setminus b_K) \geq (1-\epsilon)^2\mathfrak{M}(\mathbb{R}^n)$. Let us choose (using step 1) a large K and a small $\delta_2 > 0$, such that $\mathfrak{M}(b_{K/2}(\delta_1/2, 2\delta_2)) > (1-\epsilon)\mathfrak{M}(\mathbb{R}^n)$ for any $\delta_1 > 0$. Then, for any fixed $\delta_1 > 0$, for all large r, $\mathfrak{M}^r(b_K(\delta_1, \delta_2)) > (1-\epsilon)\mathfrak{M}(\mathbb{R}^n)$. Now, using Lemma 2.59(ii), we can choose K' and T_K sufficiently large, and then δ_1 sufficiently small, so that, uniformly on the initial states $\hat{\Psi}^r_{\mathcal{E}}(0) \in b_K(\delta_1, \delta_2)$,

$$\liminf_{r\to\infty} \mathbb{P}\{\hat{\Psi}^r_{\mathcal{E}}(T_K) \in b_{K'} \setminus b_{2K}\} \geq 1 - \epsilon.$$

Therefore,

$$\mathfrak{M}^r(b_{K'} \setminus b_{2K}) > (1-\epsilon)^2\mathfrak{M}(\mathbb{R}^n)$$

for all large r, and then for the limiting measure $\mathfrak{M}$ we must have $\mathfrak{M}(\mathbb{R}^n \setminus b_K) \geq (1-\epsilon)^2\mathfrak{M}(\mathbb{R}^n)$. □

7.3. Steady state behaviour of locally unstable systems: evanescence of invariant measures in Halfin-Whitt regime. In this section we show that for any system satisfying the conditions of Proposition 2.44, considered in the Halfin-Whitt asymptotic regime, the stationary distributions of $\hat{X}^r_{\mathcal{I}}(\cdot)$ and of $\hat{\Psi}^r_{\mathcal{E}}(\cdot)$ escape to infinity as $r \to \infty$. We will rely on the results of Theorem 2.57.

THEOREM 2.60. *Consider a set of parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$ satisfying Proposition 2.44. Consider a sequence of systems in the Halfin-Whitt asymptotic regime (Definition 2.55). Denote by $\mathfrak{M}^r$ the stationary distribution of the process $\hat{X}^r_{\mathcal{I}}(\cdot)$, a probability measure on $\mathbb{R}^I$. Let $b_K = \{z : |z| \leq K\} \subset \mathbb{R}^I$ be the ball of radius K in $\mathbb{R}^I$. Then for any K, $\mathfrak{M}^r(b_K) \to 0$ as $r \to \infty$.*

PROOF. We will show the result for the projection $\pi\hat{X}_{\mathcal{I}}$ of the limiting state $\hat{X}_{\mathcal{I}}$ onto the subspace $\mathcal{N} = \{z \in \mathbb{R}^I : \sum z_i = 0\}$. Since $(1, \dots, 1)^\top$ is an eigenvector of A_u with a real eigenvalue, we have

$$\pi A_u y = \pi A_u \pi y$$

for all $y \in \mathbb{R}^I$. (Recall $A_c = \pi A_u$ as transformations by Lemma 2.24.) Therefore, for the map F defined in (32), we have

$$\pi A_u F(y) = \pi A_u \pi y = A_c \pi y$$

for all $y \in \mathbb{R}^I$.

From this and Theorem 2.57 we see that $\pi\hat{X}_{\mathcal{I}}$ satisfies the linear SDE

$$\pi\hat{X}_{\mathcal{I}}(t) = \pi\hat{X}_{\mathcal{I}}(0) + \int_0^t A_c(\pi\hat{X}_{\mathcal{I}}(s))ds + \pi\left(\sqrt{2\lambda_i}B_i(t)\right)_{i\in\mathcal{I}}. \tag{38}$$

The argument now proceeds as in the proof of Theorem 2.58, using the unstable ODE

$$\dot{z} = (MA_cM')z, z \in \mathcal{M} \cap \mathcal{N}$$

in the place of (37). (The entire analysis will proceed in the $(I-1)$-dimensional space $\mathcal{N}$.) □

We conjecture that a stronger result holds:

Conjecture 2.61. Consider a set of parameters $\lambda_{\mathcal{I}}$, $\beta_{\mathcal{J}}$, $\mu_{\mathcal{E}}$ satisfying the conditions of Theorem 2.58, respectively 2.60. Consider a sequence of systems in underload, respectively the Halfin-Whitt asymptotic regime. For $0 \le \epsilon < \frac{1}{2}$, denote by $\mathfrak{M}^r$ the stationary distribution of the process $r^{-\frac{1}{2}+\epsilon}(X^r_{\mathcal{I}}(\cdot) - rx^*_{\mathcal{I}})$, a probability measure on $\mathbb{R}^I$. Let $b_K = \{z : |z| \le K\} \subset \mathbb{R}^I$ be the ball of radius K in $\mathbb{R}^I$. Then for any K, $\mathfrak{M}^r(b_K) \to 0$ as $r \to \infty$.

That is, we suspect that the limiting measure is non-tight on all scales strictly smaller than fluid (corresponding to $\epsilon = \frac{1}{2}$ above).

7.4. Diffusion scale tightness of stationary distributions for the case when service rate depends on the server type only. In this section we consider a special case when there exists a set of positive rates $\mu_{\mathcal{J}}$, such that $\mu_{ij} = \mu_j$ for all $(ij) \in \mathcal{E}$. We demonstrate tightness of invariant distributions of the diffusion-scaled process, assuming the system is critically loaded on the fluid scale, i.e. $\rho = 1$. (An analogous result holds for the underloaded system, but critical load is typically more relevant in applications.) This, in combination with the transient diffusion limit results, allows us to claim that the limit of invariant distributions is the invariant distribution of the limiting diffusion process.

We will work in the Halfin-Whitt asymptotic regime specified by Definition 2.55. As noted after the definition, we use $\hat{Q}^r_i(t) = Q^r_i(t)/\sqrt{r}$, $Z^r_j(t) = \Psi^r_j(t) - r\beta_j$, $\hat{Z}^r_j(t) = Z^r_j(t)/\sqrt{r}$; here, $\hat{Z}^r_j(t)$ measures the deviation of pool-j occupancy from full occupancy $r\beta_j$, rather than from the equilibrium value for the rth network, $\rho^r r\beta_j$. Our choice of signs is such that $\hat{Q}^r_i \ge 0$ while $\hat{Z}^r_j \le 0$.

Theorem 2.62. *Suppose $\mu_{ij} = \mu_j$, $(ij) \in \mathcal{E}$ and $\rho = 1$. Consider a system under the LQFS-LB rule in the Halfin-Whitt asymptotic regime (Definition 2.55). Then, for any real*

$$\theta < \theta_0 := \frac{2\min_i \lambda_i}{\sum_i \lambda_i + (\max_j \mu_j)\sum_j \beta_j},$$

the stationary distributions are such that

$$\limsup_r \mathbb{E}\left[\sum_i \exp(\theta\hat{Q}^r_i) + \sum_j \beta_j \exp(\theta\hat{Z}^r_j/\beta_j)\right] < \infty.$$

Proof. Note that the statement is trivial for $\theta = 0$. Also, for $\theta > 0$ each term $\exp(\theta\hat{Z}^r_j/\beta_j)$ is bounded so has finite expectation, while for $\theta < 0$ each term $\exp(\theta\hat{Q}^r_i)$ is bounded so has finite expectation.

Our method is based on that in [Gamarnik and Stolyar, 2012]. (The exposition below is self-contained.)

Step 1: preliminary bounds. Consider the embedded Markov chain taken at the instants of (say, right after) the transitions. We will use uniformisation. That is, we keep the total rate of all transitions from any state constant at

$$\alpha^r r = \sum_i \lambda_i^r + \sum_j r\beta_j\mu^*, \quad \mu^* \equiv \max \mu_j;$$

note that, as $r \to \infty$, $\alpha^r \to \alpha^* = \sum_i \lambda_i + \sum_j \beta_j\mu^*$[14]. The transitions are of three types: arrivals, departures, and virtual transitions, which do not change the state of the system. The rate of a transition due to a type i arrival is λ_i^r; for the service completion at pool j the rate is $\mu_j(r\beta_j + Z_j^r)$ (recall $Z_j^r \le 0$); and a virtual transition occurs at the complementary rate $\alpha^r r - \sum_i \lambda_i^r - \sum_j \mu_j(r\beta_j + Z_j^r)$. (The probability that a transition occurring at a transition instant has a given type is the ratio of the corresponding rate and $\alpha^r r$.) The stationary distribution of the embedded, uniformised Markov chain is the same as that of the original, continuous-time chain.

In the rest of the proof, $\tau \in \{0, 1, 2, \ldots\}$ refers to the discrete time of the embedded Markov chain.

We will work with the following Lyapunov function

$$\mathcal{L}(\tau) := \sum_i \exp(\theta\hat{Q}_i^r(\tau)) + \sum_j \beta_j \exp(\theta\hat{Z}_j^r(\tau)/\beta_j). \tag{39}$$

Throughout, we use the bound

$$\exp(\theta y) \le \exp(\theta x)\left(1 + \theta(y-x) + \frac{1}{2}\theta^2(y-x)^2\exp(\theta\,|y-x|)\right) \tag{40}$$

which arises from the second-order Taylor expansion of $\exp(\theta y)$.

A priori we do not know that $\mathbb{E}[\mathcal{L}(\tau)]$ exists for $\theta > 0$. Indeed, while $\hat{Z}_j^r(t)$ is bounded for any r (above by 0 and below by $-\beta_j\sqrt{r}$), the scaled queue size $\hat{Q}_i^r(t)$ is unbounded. To deal with this, we also consider the truncated Lyapunov function $\mathcal{L}^K = \min\{\mathcal{L}, K\}$.

In the equation below, let x denote the variable of interest (either $\hat{Q}_i^r$ or $\hat{Z}_j^r/\beta_j$), and let $S(\tau)$ denote the state of the embedded Markov chain at time τ. From (40) we obtain

$$\begin{aligned}\mathbb{E}[\exp(\theta x(\tau+1)) - \exp(\theta x(\tau)) \mid S(\tau)] \le \\ \exp(\theta x(\tau))\Big(\theta\mathbb{E}[x(\tau+1) - x(\tau) \mid S(\tau)] + \\ \frac{1}{2}\theta^2\mathbb{E}\left[(x(\tau+1)-x(\tau))^2\exp(\theta\,|x(\tau+1)-x(\tau)|) \mid S(\tau)\right]\Big).\end{aligned}$$

Since for both $\hat{Z}_j^r$ and $\hat{Q}_i^r$ the change in a single transition is bounded by $1/\sqrt{r}$, we conclude:

$$\begin{aligned}\mathbb{E}[\exp(\theta\hat{Q}_i^r(\tau+1)) - \exp(\theta\hat{Q}_i^r(\tau))|S(\tau)] \le \\ \exp(\theta\hat{Q}_i^r(\tau))\left(\theta\mathbb{E}[\hat{Q}_i^r(\tau+1) - \hat{Q}_i^r(\tau)|S(\tau)] + \left(\frac{1}{2}\theta^2\exp(\theta/\sqrt{r})\right)\frac{1}{r}\right),\end{aligned} \tag{41a}$$

$$\begin{aligned}\mathbb{E}[\beta_j\exp(\theta\hat{Z}_j^r(\tau+1)/\beta_j) - \beta_j\exp(\theta\hat{Z}_j^r(\tau)/\beta_j)|S(\tau)] \le \\ \exp(\theta\hat{Z}_j^r(\tau)/\beta_j)\Big(\theta\mathbb{E}[\hat{Z}_j^r(\tau+1) - \hat{Z}_j^r(\tau)|S(\tau)] + \left(\frac{1}{\beta_j}\frac{1}{2}\theta^2\exp(\theta/\sqrt{r})\right)\frac{1}{r}\Big).\end{aligned} \tag{41b}$$

Clearly, as long as values of θ are bounded, for any fixed $C_2 > 1$ and all sufficiently (depending on C_2) large r, the second summands in (41a) and (41b) are bounded above

[14]This use of α^r and α^* is unrelated to the rate or processing workload α_j defined in (2).

by $C_2\frac{1}{2}\theta^2\frac{1}{r}$ and $\frac{1}{\beta_*}C_2\frac{1}{2}\theta^2\frac{1}{r}$, respectively, where $\beta_* = \min_j \beta_j$. Note that the second bound is independent of j. That is, we obtain

$$\text{(42a)}\quad \mathbb{E}[\exp(\theta\hat{Q}_i^r(\tau+1)) - \exp(\theta\hat{Q}_i^r(\tau))|S(\tau)] \le \exp(\theta\hat{Q}_i^r(\tau))\left(\theta\mathbb{E}[\hat{Q}_i^r(\tau+1) - \hat{Q}_i^r(\tau)|S(\tau)] + C_2\frac{1}{2}\theta^2\frac{1}{r}\right)$$

$$\text{(42b)}\quad \mathbb{E}[\beta_j\exp(\theta\hat{Z}_j^r(\tau+1)/\beta_j) - \beta_j\exp(\theta\hat{Z}_j^r(\tau)/\beta_j)|S(\tau)] \le \exp(\theta\hat{Z}_j^r(\tau)/\beta_j)\left(\theta\mathbb{E}[\hat{Z}_j^r(\tau+1) - \hat{Z}_j^r(\tau)|S(\tau)] + \frac{1}{\beta_*}C_2\frac{1}{2}\theta^2\frac{1}{r}\right)$$

Next, we will obtain an upper bound on the drift

$$\mathbb{E}[\mathcal{L}(\tau+1) - \mathcal{L}(\tau)|S(\tau)].$$

To do that, we introduce an artificial scheduling/routing rule, which acts only within one time step, and is such that the increment $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ under this rule is "almost" a (pathwise, w.p.1) upper bound on this increment under the actual – LQFS-LB – rule. (It is important to keep in mind that the artificial rule is *not* a rule that is applied continuously. It is limited to one time step, and its sole purpose is to derive a pathwise upper bound on the increment $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ within one time step.)

Step 2: Artificial scheduling/routing rule. We will use the following notation:

$$\mathcal{I}_+ = \mathcal{I}_+(\tau) \equiv \{i : \hat{Q}_i^r(\tau) > 0\}, \qquad \mathcal{I}_0 = \mathcal{I}_0(\tau) \equiv \{i : \hat{Q}_i^r(\tau) = 0\},$$
$$\mathcal{J}_- = \mathcal{J}_-(\tau) \equiv \{j : \hat{Z}_j^r(\tau) < 0\}, \qquad \mathcal{J}_0 = \mathcal{J}_0(\tau) \equiv \{j : \hat{Z}_j^r(\tau) = 0\}$$

Artificial scheduling: Departures from servers $j \in \mathcal{J}_-$ are processed normally, i.e. reduce the corresponding $Z_j^r(\tau)$ by 1. Whenever there is a departure from a server pool $j \in \mathcal{J}_0$, the server picks a customer type i with nominal probability $\lambda_{ij}^r/\sum_i \lambda_{ij}^r$. If the chosen i is one of the types in $\mathcal{I}_+$, then we keep $Z_j^r(\tau+1) = 0$ and reduce $Q_i^r(\tau+1) = Q_i^r(\tau) - 1$. However, if $i \in \mathcal{I}_0$, i.e. $Q_i^r(\tau) = 0$, then we keep $Q_i^r(\tau+1) = Q_i^r(\tau) = 0$ and instead allow $Z_j^r(\tau+1) = -1$.

Artificial routing: Arrivals to customer types $i \in \mathcal{I}_+$ are processed normally, i.e. increase the corresponding $Q_i^r(\tau)$ by 1. Whenever there is an arrival to a customer type $i \in \mathcal{I}_0$, the customer picks a server type j with nominal probability $\lambda_{ij}^r/\lambda_i^r$. If the chosen j is one of the types in $\mathcal{J}_-$, then we keep $Q_i^r(\tau+1) = Q_i^r(\tau) = 0$ and increase $Z_j^r(\tau+1) = Z_j^r(\tau) + 1$. However, if the chosen $j \in \mathcal{J}_0$, i.e. $Z_j^r(\tau) = 0$, then we keep $Z_j^r(\tau+1) = Z_j^r(\tau) = 0$ and instead allow $Q_i^r(\tau+1) = 1$.

Step 3: One time-step drift under the artificial rule. For $i \in \mathcal{I}_+$,

$$\mathbb{E}[\hat{Q}_i^r(\tau+1) - \hat{Q}_i^r(\tau)|S(\tau)] = \frac{1}{\alpha^r r}\frac{1}{\sqrt{r}}\left(\lambda_i^r - \sum_j(\mu_j r\beta_j)\frac{\lambda_{ij}^r}{\sum_k \lambda_{kj}^r}\right).$$

Recalling that

$$\sum_k \lambda_{kj}^r = \mu_j\beta_j r\rho^r = \mu_j\beta_j r(1 - C/\sqrt{r}), \tag{43}$$

we obtain

$$\mathbb{E}[\hat{Q}_i^r(\tau+1) - \hat{Q}_i^r(\tau)|S(\tau)] = -\frac{C\lambda_i}{\alpha^*}\frac{1+o(1)}{r},\ i \in \mathcal{I}_+, \tag{44}$$

where $o(1)$ is a fixed function, vanishing as $r \to \infty$.

If $\hat{Q}_i^r(\tau) = 0$ (i.e. $i \in \mathcal{I}_0$), and a new arrival of type i picks a server pool j which has idle servers, (i.e. $j \in \mathcal{J}_-$), then $\hat{Q}_i^r$ stays at 0 and $\hat{Q}_i^r(\tau+1) - \hat{Q}_i^r(\tau) = 0$. However, if a new type i arrival picks some server pool $j \in \mathcal{J}_0$ which has no available idle servers, then (by the definition of artificial rule) $\hat{Q}_i^r(\tau+1) - \hat{Q}_i^r(\tau) = \hat{Q}_i^r(\tau+1) = 1/\sqrt{r}$. Thus, we can write:

$$\mathbb{E}[\hat{Q}_i^r(\tau+1) - \hat{Q}_i^r(\tau)|S(\tau)] = \sum_{j\in\mathcal{J}_0} \frac{\lambda_{ij}^r}{\alpha^r r}\frac{1}{\sqrt{r}},\ i \in \mathcal{I}_0. \tag{45}$$

The right-hand side of (45) is of order $1/\sqrt{r}$. This may be alarming, because the time step is of order $\frac{1}{r}$, and we would like to avoid large jumps in a single time step. However, we will see shortly that order $1/\sqrt{r}$ terms in $\mathbb{E}[\mathcal{L}(\tau+1) - \mathcal{L}(\tau)|S(\tau)]$ cancel out, and the expected drift is in fact of order $1/r$ (same order of magnitude as the time step).

The treatment of the drift of $\hat{Z}_j^r$ is similar (and again makes use of (43)). For $j \in \mathcal{J}_-$,

$$\mathbb{E}[\hat{Z}_j^r(\tau+1) - \hat{Z}_j^r(\tau)|S(\tau)] = -\frac{1}{\alpha^r}\mu_j(\hat{Z}_j^r(\tau) + \beta_j C)\frac{1}{r},\ j \in \mathcal{J}_-, \tag{46}$$

and for $j \in \mathcal{J}_0$,

$$\begin{aligned}\mathbb{E}[\hat{Z}_j^r(\tau+1) - \hat{Z}_j^r(\tau)|S(\tau)] &= -\frac{1}{\sqrt{r}}\sum_{i\in\mathcal{I}_0}\frac{r\mu_j\beta_j}{\alpha^r r}\frac{\lambda_{ij}^r}{\sum_k \lambda_{kj}^r} \\ &= -\frac{1}{1 - C/\sqrt{r}}\sum_{i\in\mathcal{I}_0}\frac{\lambda_{ij}^r}{\alpha^r r}\frac{1}{\sqrt{r}},\ j \in \mathcal{J}_0.\end{aligned} \tag{47}$$

We can rewrite (47) as

$$\mathbb{E}[\hat{Z}_j^r(\tau+1) - \hat{Z}_j^r(\tau)|S(\tau)] = -\sum_{i\in\mathcal{I}_0}\frac{\lambda_{ij}^r}{\alpha^r r}\frac{1}{\sqrt{r}} - \frac{C\sum_{i\in\mathcal{I}_0}\lambda_{ij}}{\alpha^*}\frac{1+o(1)}{r},\ j \in \mathcal{J}_0, \tag{48}$$

where $o(1)$ is a fixed function, vanishing as $r \to \infty$.

Note that if $\mathcal{L}(\tau) \geq K$ then $\mathcal{L}^K$ cannot increase over the next time step. The drift of $\mathcal{L}^K(\tau)$ starting from a value $\mathcal{L}(\tau) < K$ is no greater than if we allowed the transitions that increase $\mathcal{L}(\tau)$ above K. Putting together this observation and equations (42), (44) – (48), we obtain

$$\text{(49a)} \qquad \mathbb{E}[\mathcal{L}^K(\tau+1) - \mathcal{L}^K(\tau) | S(\tau)] \leq$$

$$\text{(49b)} \qquad \mathbf{1}_{\{\mathcal{L}(\tau)\leq K\}} \left(\sum_{i\in\mathcal{I}_+} \exp(\theta \hat{Q}_i^r(\tau)) \theta \Big(-\frac{C\lambda_i(1+o(1))}{\alpha^*} \Big) \frac{1}{r} \right.$$

$$\text{(49c)} \qquad + \sum_{i\in\mathcal{I}_0, j\in\mathcal{J}_0} \theta \lambda_{ij}^r \frac{1}{\alpha^r r} \frac{1}{\sqrt{r}}$$

$$\text{(49d)} \qquad + \sum_{j\in\mathcal{J}_-} \exp(\theta \hat{Z}_j^r(\tau)/\beta_j) \theta \Big(-\frac{\mu_j}{\alpha^r} \Big) \big(\hat{Z}_j^r(\tau) + \beta_j C \big) \frac{1}{r}$$

$$\text{(49e)} \qquad + \sum_{j\in\mathcal{J}_0, i\in\mathcal{I}_0} \theta \Big(-\lambda_{ij}^r \frac{1}{\alpha^r r} \frac{1}{\sqrt{r}} - \frac{C\lambda_i(1+o(1))}{\alpha^*} \frac{1}{r} \Big)$$

$$\text{(49f)} \qquad + \sum_{i\in\mathcal{I}} \exp(\theta \hat{Q}_i^r(\tau)) \Big(\frac{C_2}{2} \theta^2 \Big) \frac{1}{r}$$

$$\text{(49g)} \qquad \left. + \sum_{j\in\mathcal{J}} \frac{1}{\beta_*} \exp(\theta \hat{Z}_j^r(\tau)/\beta_j) \Big(\frac{C_2}{2} \theta^2 \Big) \frac{1}{r} \right).$$

(There is no exponential in (49c) and (49e) because by assumption the relevant queues, respectively idlenesses, are equal to zero.) Note that the $O(1/\sqrt{r})$ terms in (49c) and (49e) cancel each other as promised, so there will be no $O(1/\sqrt{r})$ terms in the final bound. We will show that this bound is in fact negative later, in Step 5.

Step 4: One time-step drift under the LQFS-LB rule. We now explain in what sense the increment $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ under the artificial rule is "almost" an upper bound on this increment under LQFS-LB. To illustrate the idea, suppose first that all β_j are equal. Then, as we will now show, the routing or scheduling decision made by LQFS-LB at time step τ will have a smaller increment $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ than the artificial rule (with probability 1). Suppose the decision is associated with scheduling a customer from a queue after a service completion at server $j \in \mathcal{J}_0$. (After service completion at a server $j \in \mathcal{J}_-$ the two rules behave identically.) Suppose first that LQFS-LB schedules a customer from queue i, while the artificial policy attempts to schedule a customer from queue i'. Then by definition of LQFS-LB, $\hat{Q}_i^r \geq \hat{Q}_{i'}^r$, so the one-step increment $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ is smaller for LQFS-LB. If the artificial rule chooses i' with $\hat{Q}_{i'}^r = 0$, then LQFS-LB will decrease $\hat{Q}_i^r$ while the artificial rule increases $\hat{Z}_j^r$. Convexity of the exponential function shows that in this case, the one-step increment $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ is again smaller for LQFS-LB. We argue similarly when the decision to be taken by the rules is the routing of a newly arrived customer of type i. Therefore, when all β_j are equal, the key estimate (49) of the expected drift holds, in exactly same form, for the LQFS-LB rule as well.

Now consider the case of general β_j. In the event of a service completion (and then possibly taking a customer for service from one of the non-zero queues), the increment $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ under LQFS-LB is still clearly no greater than under the artificial rule. The only situation when LQFS-LB can possibly cause a greater increment than the artificial rule is as follows. There is an arrival of a type i customer, which the artificial rule routes to pool j with $\hat{Z}_j^r < 0$, but the LQFS-LB will instead route it to pool k such that $\hat{Z}_j^r/\beta_j \geq \hat{Z}_k^r/\beta_k$. Given convexity of the function $e^{\theta x}$, the largest increment of $\mathcal{L}(\tau+1) - \mathcal{L}(\tau)$ occurs when $\hat{Z}_j^r/\beta_j = \hat{Z}_k^r/\beta_k$. (If $\theta > 0$, increasing $\hat{Z}_k^r$ would make the positive increment larger; if $\theta < 0$, increasing $\hat{Z}_k^r$ would make the negative increment get smaller in absolute value.)

Thus, as we replace the artificial rule by LQFS-LB, in the "worst case", the increment

$$\beta_j \exp(\theta[\hat{Z}_j^r(\tau) + r^{-1/2}]/\beta_j) - \beta_j \exp(\theta \hat{Z}_j^r(\tau)/\beta_j)$$

may need to be replaced by

$$\beta_k \exp(\theta[\hat{Z}_k^r(\tau) + r^{-1/2}]/\beta_k) - \beta_k \exp(\theta \hat{Z}_k^r(\tau)/\beta_k),$$

with $\hat{Z}_k^r(\tau)$ satisfying $\hat{Z}_j^r(\tau)/\beta_j = \hat{Z}_k^r(\tau)/\beta_k$. In this case, after applying (40) we obtain

$$\beta_k \exp(\theta \hat{Z}_k^r(\tau+1)/\beta_k) - \beta_k \exp(\theta \hat{Z}_k^r(\tau)/\beta_k) \leq \exp(\theta \hat{Z}_k^r(\tau)/\beta_k) \left(\theta r^{-1/2} + \left(\frac{1}{\beta_k} \frac{1}{2} \theta^2 \exp(\theta/\sqrt{r}) \right) \frac{1}{r} \right),$$

which is bounded above by

$$\exp(\theta \hat{Z}_j^r(\tau)/\beta_j) \left(\theta r^{-1/2} + \left(\frac{1}{\beta_*} \frac{1}{2} \theta^2 \exp(\theta/\sqrt{r}) \right) \frac{1}{r} \right)$$

(where we have used $\hat{Z}_k^r(\tau)/\beta_k = \hat{Z}_j^r(\tau)/\beta_j$). Thus, (42b) remains true even if we use LQFS-LB rather than the artificial rule; so the estimate (49) continues to hold.

Step 5: Exponential moments estimates. Next, note that for each fixed $K > 0$ and each fixed parameter r, the values of $\exp(\theta \hat{Q}_i^r(\tau))$ are uniformly bounded over all states $S(\tau)$ satisfying the condition
$\mathcal{L}(\tau) \leq K$; the values of $\exp(\theta \hat{Z}_j^r(\tau)/\beta_j)$ are "automatically" uniformly bounded (for a fixed r). We take the expected values of both parts of (49) with respect to the invariant distribution. The expectation of the left-hand side is of course 0, and so we get rid of the factor $1/r$ from the right-hand side expectation. The resulting estimates we will write separately for the cases $\theta > 0$ and $\theta < 0$ (with the case $\theta = 0$ being trivial).

Case $\theta > 0$. For a fixed $\theta > 0$, the expected value of the sum of all terms not containing $\exp(\theta \hat{Q}_i^r(\tau))$ is bounded (uniformly in r). Indeed, this follows from the facts that $\hat{Z}_j^r(\tau) \leq 0$ and

$$0 \leq -\theta \hat{Z}_j^r(\tau) \exp(\theta \hat{Z}_j^r(\tau)/\beta_j) \leq \beta_j/e$$

(because $0 \geq xe^x \geq -\frac{1}{e}$ for $x \leq 0$). Then, we obtain:

$$\mathbb{E}\left[\mathbf{1}_{\{\mathcal{L}(\tau) \leq K\}} \sum_{i \in \mathcal{I}_+} \exp(\theta \hat{Q}_i^r(\tau)) \left(\frac{C \lambda_i (1 + o(1))}{\alpha^*} \theta - \left(\frac{C_2}{2} \theta^2 \right) \right) \right] \leq C_1 \tag{50}$$

for some constant $C_1 = C_1(\theta) > 0$, uniformly on all sufficiently large r. Now let us fix a sufficiently small positive θ, so that all coefficients of $\exp(\theta \hat{Q}_i^r(\tau))$ are at least some $\epsilon > 0$ (for all large r). Recalling that $C_2 > 1$ can be arbitrarily close to 1, it suffices that $\theta < \theta_0 = 2(\min_i \lambda_i)/\alpha^*$. Then,

$$\mathbb{E}\left[\mathbf{1}_{\{\mathcal{L}(\tau) \leq K\}} \sum_{i \in \mathcal{I}_+} \exp(\theta \hat{Q}_i^r(\tau)) \right] \leq C_1/\epsilon,$$

from where, letting $K \to \infty$, by monotone convergence, we obtain

$$\mathbb{E}\left[\sum_{i \in \mathcal{I}_+} \exp(\theta \hat{Q}_i^r(\tau)) \right] \leq C_1/\epsilon < \infty, \tag{51}$$

uniformly on all large r.

Case $\theta < 0$. Fix arbitrary $\theta < 0$. In this case, the expected value of the sum of all terms not containing $\exp(\theta \hat{Z}_j^r(\tau))$ is bounded (uniformly on r). We can write:

$$\mathbb{E}\left[\mathbf{1}_{\{\mathcal{L}(\tau)\le K\}} \sum_{j\in\mathcal{J}_-} \exp(\theta \hat{Z}_j^r(\tau)/\beta_j)\left(\theta[\frac{\mu_j}{\alpha^r}][\hat{Z}_j^r(\tau)+\beta_j C] - \left(\frac{1}{\beta_*}\frac{C_2}{2}\theta^2\right)\right)\right] \le C_1', \tag{52}$$

for some constant $C_1' = C_1'(\theta) > 0$, uniformly on all sufficiently large r. Let us choose sufficiently large $K_1 > 0$, such that the condition $\hat{Z}_j^r(\tau) \le -K_1$ implies that

$$\left(\theta\left(\frac{\mu_j}{\alpha^r}\right)\left(\hat{Z}_j^r(\tau)+\beta_j C\right) - \left(\frac{1}{\beta_*}\frac{C_2}{2}\theta^2\right)\right) \ge \epsilon,$$

for some $\epsilon > 0$ (and all large r). Then, from (52),

$$\mathbb{E}\left[\mathbf{1}_{\{\mathcal{L}(\tau)\le K\}} \sum_{j\in\mathcal{J}_-} \mathbf{1}_{\{\hat{Z}_j^r(\tau)\le -K_1\}} \exp(\theta \hat{Z}_j^r(\tau)/\beta_j)\right] \le C_1'/\epsilon,$$

from where, letting $K \to \infty$, by monotone convergence, we obtain

$$\mathbb{E}\left[\sum_{j\in\mathcal{J}_-} \mathbf{1}_{\{\hat{Z}_j^r(\tau)\le -K_1\}} \exp(\theta \hat{Z}_j^r(\tau)/\beta_j)\right] \le C_1'/\epsilon < \infty,$$

uniformly on all large r, which implies the required result. □

Corollary 2.63. *The sequence of stationary distributions of the processes* $\left(\hat{Q}_{\mathcal{I}}^r(\cdot), \hat{Z}_{\mathcal{J}}^r(\cdot)\right)$ *has a weak limit, which is the unique stationary distribution of the limiting process* $\left(\hat{Q}_{\mathcal{I}}(\cdot), \hat{Z}_{\mathcal{J}}(\cdot)\right)$*, described as follows:*

$$\hat{Q}_i(t) \equiv \max\{\hat{Y}(t)/I, 0\},\ \forall i,\ \ \hat{Z}_j(t) \equiv \min\left\{\frac{\beta_j}{\sum_k \beta_k}\hat{Y}(t), 0\right\},\ \forall j, \tag{53}$$

where $\hat{Y}(\cdot)$ *is a one-dimensional diffusion process with constant variance parameter* $2\sum_i \lambda_i$ *and piecewise linear drift, equal at point* x *to*

$$-\left(\sum_j \mu_j\right)(C + \min\{x, 0\}).$$

The invariant distribution density is then a continuous function, which is a "concatenation" at point 0 *of exponential (for* $x \ge 0$*) and Gaussian (for* $x \le 0$*) distribution densities.*

Proof. Theorem 2.62 implies tightness of stationary distributions of $\left(\hat{Q}_{\mathcal{I}}^r(\cdot), \hat{Z}_{\mathcal{J}}^r(\cdot)\right)$. Then, it follows from [Liptser and Shiryaev, 1989, Theorem 8.5.1] (whose conditions are easily verified in our case), that as $r \to \infty$, any weak limit of the sequence of stationary distributions of the processes $\left(\hat{Q}_{\mathcal{I}}^r(\cdot), \hat{Z}_{\mathcal{J}}^r(\cdot)\right)$ is a stationary distribution of the limit process. This limiting process is the one-dimensional diffusion given by (53) (see [Gurvich and Whitt, 2009, Theorem 4.4]), and it is easy to see that its invariant distribution is the "concatenation" specified above. □

A tightness result analogous to Theorem 2.62 also holds for the underloaded system, $\rho < 1$, and can be proved by a similar method. The asymptotic regime in this case is such that $\lambda_i^r = r\lambda_i$ (there is no point in considering $O(\sqrt{r})$ terms in λ_i^r when $\rho < 1$). We denote $Z_j^r(t) = \Psi_j^r(t) - r\beta_j\rho$ (which is consistent with the definition given earlier in

this section for $\rho = 1$), and keep notation $Q_i^r(t)$ for the queue length. We work with the following Lyapunov function:

$$\mathcal{L} \equiv \sum_i \left[\exp(\theta(1-\rho)\sqrt{r} + \theta\hat{Q}_i^r) - \exp(\theta(1-\rho)\sqrt{r})\right] + \sum_j \beta_j \exp(\theta\hat{Z}_j^r/\beta_j).$$

The same approach as in the proof of Theorem 2.62 leads to the following result: for any real θ,

$$\limsup_r \mathbb{E}[\sum_j \exp(\theta\hat{Z}_j^r)] < \infty.$$

The limiting process for $\hat{Z}_{\mathcal{J}}^r(\cdot)$ is $\hat{Z}_{\mathcal{J}}(\cdot) \equiv (\frac{\beta_j}{\sum_k \beta_k}\hat{Y}(\cdot))$, with $\hat{Y}(\cdot)$ being a one-dimensional Ornstein-Uhlenbeck process, with Gaussian stationary distribution. The limit of stationary distributions of $\hat{Z}_{\mathcal{J}}^r(\cdot)$ is the (Gaussian) stationary distribution of $\hat{Z}_{\mathcal{J}}(\cdot)$.

8. LAP steady-state on sub-fluid scales

8.1. Main theorem and set-up. The main result of this section is to show that not only is LAP stable on the fluid scale, it is in fact stable on essentially all scales larger than the diffusion scale.

THEOREM 2.64. *Consider the sequence of systems under LAP policy, in the scaling regime and under the assumptions specified in §1, with $\rho < 1$. Then:*

1. *For all sufficiently large r, the system is stable, i.e. the countable state-space Markov chain $(\Psi_{\mathcal{E}}^r(\cdot), Q_{\mathcal{I}}^r(\cdot))$ is positive recurrent.*
2. *For any $\epsilon > 0$, the stationary distribution of $r^{-1/2-\epsilon}(\Psi_{\mathcal{E}}^r(\cdot) - r\psi_{\mathcal{E}}^*, Q_{\mathcal{I}}^r(\cdot))$ weakly converges to 0.*

Theorem 2.51 proves statement (1), so for all large r we may define steady-state variables $(\Psi_{\mathcal{E}}^r, Q_{\mathcal{I}}^r)$ s.t. $\Psi_{\mathcal{E}}^r(t) \xrightarrow{w} \Psi_{\mathcal{E}}^r$, $Q_{\mathcal{I}}^r(t) \xrightarrow{w} Q_{\mathcal{I}}^r$. Moreover, Theorem 2.51 implies statement (2) for $\epsilon = 1/2$; that is,

$$\lim_{r\to\infty} \mathbb{P}\left(\left\|\frac{1}{r}(\Psi_{\mathcal{E}}^r(\cdot) - \psi_{\mathcal{E}}^* r, Q_{\mathcal{I}}^r(\cdot))\right\| > \delta\right) = 0, \tag{54}$$

for any $\delta > 0$. (The theorem statement is about weak convergence, but weak convergence to a constant implies convergence in probability.) The rest of this section is devoted to extending this result to all $\epsilon > 0$. This will involve studying finer rescalings of the process, which we call the *hydrodynamic* and *local-fluid* scalings.

Fix ϵ, $0 < \epsilon < 1/2$.

From (54), for an arbitrarily small fixed $\delta > 0$, we can choose a positive function $g(r) = o(r)$, such that

$$\mathbb{P}\{\|(\Psi_{\mathcal{E}}^r - r\psi_{\mathcal{E}}^*, Q_{\mathcal{I}}^r)\| \le g(r)\} \ge 1 - \delta. \tag{55}$$

Without loss of generality, assume $r^{-1/2-\epsilon}g(r) \to \infty$.

We will prove that there exist positive constants C and T, such that for any fixed $\delta_1 > 0$ the following holds for all sufficiently large r:

$$r^{1/2+\epsilon} \le \left\|\left(\Psi_{\mathcal{E}}^r(0) - r\psi_{\mathcal{E}}^*, Q_{\mathcal{I}}^r(0)\right)\right\| \le g(r) \text{ implies} \tag{56}$$
$$\mathbb{P}\left\{\left\|\left(\Psi_{\mathcal{E}}^r(T\log r) - r\psi_{\mathcal{E}}^*, Q_{\mathcal{I}}^r(T\log r)\right)\right\| \le Cr^{1/2+\epsilon}\right\} \ge 1 - \delta_1.$$

This fact, along with (55), will prove Theorem 2.64(ii).

We will need strong law of large numbers type results, which can be obtained from a strong approximation of Poisson processes, available e.g. in [Csörgő and Horváth, 1993, Chapters 1 and 2]:

PROPOSITION 2.65. *A unit rate Poisson process* $\Pi(\cdot)$ *and a standard Brownian motion* $W(\cdot)$ *can be constructed on a common probability space in such a way that the following holds. For some fixed positive constants* C_1, C_2, C_3, $\forall T > 1$ *and* $\forall u \geq 0$

$$\mathbb{P}\left(\sup_{0\leq t\leq T} |\Pi(t) - t - W(t)| \geq C_1 \log T + u\right) \leq C_2 e^{-C_3 u}.$$

Applying the result to the unit rate Poisson processes $\Pi_i^{(a)}(\cdot)$ and $\Pi_{ij}^{(s)}(\cdot)$ which drive the exogenous arrivals and departures, we obtain the following. (For $\Pi_i^{(a)}(\cdot)$, for example, we replace t with $\lambda_i r t$; T with $\lambda_i r T \log r$; and u with $r^{1/4}$.)

PROPOSITION 2.66. *For any fixed* $T > 0$ *and any subsequence of* $r \to \infty$*, we can find a further subsequence (with* r *increasing sufficiently fast), such that:*
for each $i \in \mathcal{I}$*,*

$$\sup_{0\leq t\leq T\log r} r^{-1/2-\epsilon/2} \left|\Pi_i^{(a)}(\lambda_i r t) - \lambda_i r t\right| \to 0, \ \textit{w.p.1},$$

and for each $(ij) \in \mathcal{E}$*,*

$$\sup_{0\leq t\leq T\log r} r^{-1/2-\epsilon/2} \left|\Pi_{ij}^{(s)}(\mu_{ij}\beta_j r t) - \mu_{ij}\beta_j r t\right| \to 0, \ \textit{w.p.1}.$$

Let $F^r(t)$ be the process of (unscaled) deviations from equilibrium; that is,

$$F^r(t) = (\Psi_{\mathcal{E}}^r(t) - r\psi_{\mathcal{E}}^*, Q_{\mathcal{I}}^r(t)).$$

Suppose we have a function $h(r)$, such that $r^{1/2+\epsilon} \leq h(r) \leq g(r)$. (The value $h(r)$ will be the "scale" of $F^r(0)$; sometimes, but not always, we simply use $h(r) = \|F^r(0)\|$.) We will establish properties of $F^r(\cdot)$ under two different scalings, called hydrodynamic and local-fluid.

REMARK 2.67. The use of multiple scalings (in addition to the "standard" fluid scaling) is typical in the analysis of systems in the many-server asymptotic regime, cf. [Gurvich and Whitt, 2009] and references therein. Our hydrodynamic and local-fluid scalings are somewhat unusual in that the scaling factor $h(r)$ is strictly "between" r and $r^{1/2}$. (When $h(r) = r$, both local-fluid and hydrodynamic scalings become the standard fluid scaling; if $h(r) = r^{1/2}$, the local-fluid scaling becomes the standard diffusion scaling.) Also, although the concept of analysing the system over the course of many short intervals is not new (cf. [Shah and Wischik, 2009, Section 8]), using multiple scalings simultaneously to derive tightness of stationary distributions is, to the best of our knowledge, novel.

8.2. Hydrodynamic scaling. Consider the process under the following scaling and centering:

$$\begin{aligned}(57)\quad &(\overline{\psi}_{ij}^r(t), \overline{q}_i^r(t), \overline{x}_i^r(t), \overline{a}_i^r(t), \overline{d}_{ij}^r(t), \overline{\xi}_{ij}^r(t)) = \\ &h(r)^{-1}\Big(\Psi_{ij}^r((h(r)r^{-1}t) - r\psi_{ij}^*, Q_i^r(h(r)r^{-1}t), X_i^r(h(r)r^{-1}t) - r\sum_j \psi_{ij}^*, \\ &\qquad A_i^r(h(r)r^{-1}t), D_{ij}^r(h(r)r^{-1}t), \Xi_{ij}^r(h(r)r^{-1}t)\Big)_{i\in\mathcal{I},(ij)\in\mathcal{E}}.\end{aligned}$$

Note that since $\overline{\psi}_{\mathcal{E}}^r(\cdot)$ is centered before it is scaled in space, the condition $\rho < 1$ implies $\sum_i \overline{\psi}_{ij}^r(t) \leq 0$ for all $j < J$ at all times t.

THEOREM 2.68. *Consider a sequence of deterministic realisations, such that the driving realisations satisfy the functional strong law of large numbers conditions, namely:*

$$(58)\qquad (\overline{a}_{\mathcal{I}}^r(t), \ t \geq 0) \to (\lambda_{\mathcal{I}} t, \ t \geq 0), \ \textit{u.o.c.}$$

$$(59)\quad \Big(h(r)^{-1}\big(D^r_{ij}(h(r)r^{-1}t) - \mu_{ij}\int_0^{h(r)r^{-1}t}\Psi^r_{ij}(s)ds\big),\ t\geq 0\Big)\to 0,\ \textit{u.o.c.},\ \forall (ij)\in\mathcal{E}.$$

Suppose $(\overline{\psi}^r_{\mathcal{E}}(0),\overline{q}^r_{\mathcal{I}}(0))\to(\overline{\psi}_{\mathcal{E}}(0),\overline{q}_{\mathcal{I}}(0))$. *Then, for any subsequence of* r *there exists a further subsequence along which*

$$(\overline{\psi}^r_{\mathcal{E}}(\cdot),\overline{q}^r_{\mathcal{I}}(\cdot),\overline{x}^r_{\mathcal{I}}(\cdot),\overline{a}^r_{\mathcal{I}}(\cdot),\overline{d}^r_{\mathcal{E}}(\cdot),\overline{\xi}^r_{\mathcal{E}}(\cdot))$$

converges uniformly on compact sets to a set of Lipschitz continuous functions

$$(\overline{\psi}_{\mathcal{E}}(\cdot),\overline{q}_{\mathcal{I}}(\cdot),\overline{x}_{\mathcal{I}}(\cdot),\overline{a}_{\mathcal{I}}(\cdot),\overline{d}_{\mathcal{E}}(\cdot),\overline{\xi}_{\mathcal{E}}(\cdot))$$

satisfying the hydrodynamic model equations (60). *(The conditions involving derivatives are to be satisfied whenever the derivatives exist, which is almost everywhere w.r.t. Lebesgue measure.)*

The hydrodynamic model equations are:

$$(60a)\qquad \overline{q}_i(t)\geq 0,\ \forall i\in\mathcal{I};\qquad \sum_i\overline{\psi}_{ij}(t)\leq 0,\ \forall j\in\mathcal{J}$$

$$(60b)\qquad \overline{a}_i(t)=\lambda_i t,\ \forall i\in\mathcal{I};\qquad \overline{d}_{ij}(t)=\mu_{ij}\psi^*_{ij}t,\ \forall (ij)\in\mathcal{E}$$

$$(60c)\qquad \overline{q}_i(t)=\overline{q}_i(0)+\overline{a}_i(t)-\sum_j\overline{\xi}_{ij}(t),\ \forall i\in\mathcal{I}$$

$$(60d)\qquad \overline{\psi}_{ij}(t)=\overline{\psi}_{ij}(0)+\overline{\xi}_{ij}(t)-\overline{d}_{ij}(t),\ \forall i\in\mathcal{I}$$

$$(60e)\qquad \overline{x}_i(t)=\overline{q}_i(t)+\sum_j\overline{\psi}_{ij}(t)\equiv\overline{x}_i(0),\ \forall i\in\mathcal{I}$$

$$(60f)\qquad \sum_i\overline{\psi}_{ij}(t)=0,\ \text{whenever } \overline{q}_{i'}(t)>0 \text{ for at least one } i'\in\mathcal{C}(j)$$

$$(60g)\qquad \frac{d}{dt}\overline{\xi}_{ij}(t)=0,\ \text{whenever } \overline{q}_{i'}(t)>0 \text{ for at least one } i'\in\mathcal{C}(j),\ i'<i$$

$$(60h)\qquad \frac{d}{dt}\overline{\xi}_{ij}(t)=0,\ \text{whenever } \sum_k\overline{\psi}_{kj'}(t)<0 \text{ for at least one } (ij')<(ij)$$

$$(60i)\quad \frac{d}{dt}\overline{\xi}_{ij}(t)=\min\left(\lambda_i-\sum_{(ij')<(ij)}\frac{d}{dt}\overline{\xi}_{ij'}(t),\sum_{i'}\mu_{i'j}\psi^*_{i'j}-\sum_{(i'j)<(ij)}\frac{d}{dt}\overline{\xi}_{ij'}(t)\right)$$
$$\text{whenever } \overline{q}_i(t)=0 \text{ and } \sum_k\overline{\psi}_{kj}=0.$$

Definition 2.69. We call any Lipschitz solution of (60)

$$(\overline{\psi}_{\mathcal{E}}(\cdot),\overline{q}_{\mathcal{I}}(\cdot),\overline{x}_{\mathcal{I}}(\cdot),\overline{a}_{\mathcal{I}}(\cdot),\overline{d}_{\mathcal{E}}(\cdot),\overline{\xi}_{\mathcal{E}}(\cdot))$$

a *hydrodynamic model* of the system with initial state $(\overline{\psi}_{\mathcal{E}}(0),\overline{q}_{\mathcal{I}}(0))$; a set $(\overline{\psi}_{\mathcal{E}}(\cdot),\overline{q}_{\mathcal{I}}(\cdot))$, which is a projection of a hydrodynamic model we often call a hydrodynamic model as well.

Clearly, we have the following corollary of Theorem 2.68, which we record for future reference. We denote $\overline{f}^r(\cdot)\equiv(\overline{\psi}^r_{\mathcal{E}}(\cdot),\overline{q}^r_{\mathcal{I}}(\cdot))$, $\overline{f}(\cdot)\equiv(\overline{\psi}_{\mathcal{E}}(\cdot),\overline{q}_{\mathcal{I}}(\cdot))$.

COROLLARY 2.70. *For any fixed $T > 0$, $K > 0$ and $\delta_2 > 0$, there exists a sufficiently small $\delta_3 > 0$, such that the following holds. Uniformly on all $\left\|\overline{f}^r(0)\right\| \leq K$ and all sufficiently large r, conditions*

$$\max_i \sup_{[0,T]} |\overline{a}_i^r(t) - \lambda_i t| \leq \delta_3, \tag{61}$$

$$\max_{(ij)} \sup_{[0,T]} \left| h(r)^{-1}\big(D_{ij}^r(h(r)r^{-1}t) - \mu_{ij}\int_0^{h(r)r^{-1}t} \Psi_{ij}^r(s)ds\big)\right| \leq \delta_3, \tag{62}$$

imply

$$\sup_{[0,T]} |\overline{f}^r(t) - \overline{f}(t)| \leq \delta_2, \tag{63}$$

where $\overline{f}(\cdot)$ is a hydrodynamic model with initial state $\overline{f}^r(0)$.

THEOREM 2.71. *For any $K > 0$ there exists a finite time $T = T(K)$ such that all hydrodynamic models whose starting state satisfies $\left\|(\overline{\psi}_{\mathcal{E}}(0), \overline{q}_{\mathcal{I}}(0))\right\| \leq K$ have $\sum_i \overline{\psi}_{ij}(t) = 0, \forall j < J$, $\overline{q}_i(t) = 0, \forall i \in \mathcal{I}$, and $(\overline{\psi}_{\mathcal{E}}(t), \overline{q}_{\mathcal{I}}(t)) = (\overline{\psi}_{\mathcal{E}}(T), \overline{q}_{\mathcal{I}}(T))$, for all $t \geq T$. Moreover, $(\overline{\psi}_{\mathcal{E}}(T), \overline{q}_{\mathcal{I}}(T)) = L(\overline{\psi}_{\mathcal{E}}(0), \overline{q}_{\mathcal{I}}(0))$, where L is a fixed linear mapping defined below by* (64).

PROOF. Consider the highest priority activity $(1j)$. There are two possible cases: 1 is a leaf or j is a leaf. If j is a leaf, then $\overline{\psi}_{1j}(0) \leq 0$; if the inequality is strict, $\overline{\psi}_{1j}(t)$ must increase at a positive, bounded away from 0, rate until it reaches 0 within a finite time; $\overline{\psi}_{1j}(t) = 0$ thereafter. If type 1 is a leaf, then $\overline{q}_1(t)$ must decrease and $\overline{\psi}_{1j}(t)$ increase at the same rate (positive, bounded away from 0), until the entire queue (if any) "relocates into" $\overline{\psi}_{1j}$; and after that time, $\overline{\psi}_{1j}(t)$ and $\overline{q}_1(t) = 0$ will remain constant. We see that in either case, after a finite time, the highest priority activity $(1j)$ can be in a sense ignored. This allows us to proceed by induction on the activities, from the highest priority to the lowest, to check that by some finite time T (depending on K) the hydrodynamic model gets into a state $(\overline{\psi}_{\mathcal{E}}(T), \overline{q}_{\mathcal{I}}(T))$, satisfying the conditions of the theorem, and will remain in the same state for all $t \geq T$.

Since $\overline{x}_i(t)$ do not change, the linear mapping L is as follows: $L(u_{\mathcal{E}}, w_{\mathcal{I}}) = (c_{\mathcal{E}}, 0)$ where $c_{\mathcal{E}}$ is the unique solution to

$$\sum_j u_{ij} + w_i = \sum_j c_{ij}, \ \forall i \in \mathcal{I} \tag{64a}$$

$$\sum_i c_{ij} = 0, \ \forall j < J. \tag{64b}$$

□

For future reference, note that $L(u_{\mathcal{E}}, w_{\mathcal{I}}) = (c_{\mathcal{E}}, 0)$ is a function only of the vector $z_{\mathcal{I}}$, where $z_i = w_i + \sum_j u_{ij}$. The corresponding linear mapping from $z_{\mathcal{I}}$ to $c_{\mathcal{E}}$ we denote L'.

8.3. Local-fluid scaling. The process under *local fluid scaling* is defined as follows. For each r consider

$$(\tilde{\psi}_{\mathcal{E}}^r(t), \tilde{q}_{\mathcal{I}}^r(t)) \equiv \tilde{f}^r(t) = h(r)^{-1}F^r(t).$$

We will also denote $\tilde{x}_i^r(t) = h(r)^{-1}X_i^r(t) \equiv \tilde{q}_i^r(t) + \sum_j \tilde{\psi}_{ij}^r(t)$.

Note that since $\tilde{\psi}_{\mathcal{E}}^r(\cdot)$ is centered before it is scaled in space, the condition $\rho < 1$ implies $\sum_i \tilde{\psi}_{ij}^r(t) \leq 0$ for all $j < J$ at all times t.

THEOREM 2.72. *Consider a sequence of deterministic realisations, such that the driving realisations satisfy the functional strong law of large numbers conditions, namely:*

$$(h(r)^{-1}(A^r_{\mathcal{I}}(t) - \lambda_{\mathcal{I}} rt),\ t \geq 0) \to 0,\ \textit{u.o.c.}, \tag{65}$$

$$\Big(h(r)^{-1}\big(D^r_i(t) - \mu_{ij}\int_0^t \Psi^r_{ij}(s)ds\big),\ t \geq 0\Big) \to 0,\ \textit{u.o.c.},\ \forall (ij). \tag{66}$$

Assume that the initial states converge to a fixed vector

$$(\tilde{\psi}^r_{\mathcal{E}}(0), \tilde{q}^r_{\mathcal{I}}(0)) \to (\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0)).$$

Further assume that $\tilde{q}_{\mathcal{I}}(0) = 0$ *and* $\sum_i \tilde{\psi}_{ij}(0) = 0$ *for all* $j < J$. *(In other words,* $(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0)) = L(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0))$.*) Then, for any subsequence of* r *there exists a further subsequence along which*

$$(\tilde{\psi}^r_{\mathcal{E}}(\cdot), \tilde{q}^r_{\mathcal{I}}(\cdot)) \to (\tilde{\psi}_{\mathcal{E}}(\cdot), \tilde{q}_{\mathcal{I}}(\cdot)),\ \textit{u.o.c.}, \tag{67}$$

where $(\tilde{\psi}_{\mathcal{E}}(\cdot), \tilde{q}_{\mathcal{I}}(\cdot))$ *is a set of Lipschitz functions, with initial conditions* $(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0))$, *satisfying the local fluid model equations* (69). *Moreover, these limit trajectories are such that, uniformly on all of them,*

$$\Big\|(\tilde{\psi}_{\mathcal{E}}(t), \tilde{q}_{\mathcal{I}}(t))\Big\| \leq \Big\|(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0))\Big\|\, c_1 e^{-c_2 t},\ \forall t \geq 0, \tag{68}$$

where $c_1, c_2 > 0$ *are fixed constants.*

The local fluid model equations are

$$\tilde{q}_i(t) = 0, \quad \forall i \in \mathcal{I} \tag{69a}$$

$$\sum_j \tilde{\psi}_{ij}(t) = \sum_j \tilde{\psi}_{ij}(0) - \sum_j \int_0^t \mu_{ij}\tilde{\psi}_{ij}(s)ds, \quad \forall i \in \mathcal{I} \tag{69b}$$

$$\sum_i \tilde{\psi}_{ij}(t) = 0, \quad \forall j < J \tag{69c}$$

The $I + J - 1$ equations for the $I + J - 1$ functions $(\tilde{\psi}_{ij}(\cdot))$ can be solved sequentially, in order of decreasing activity priority, since the highest unsolved-for priority will always correspond to either a customer-type or a server-type leaf of the remaining activity tree.

DEFINITION 2.73. We call any Lipschitz solution of (69)

$$(\tilde{\psi}_{\mathcal{E}}(\cdot), \tilde{q}_{\mathcal{I}}(\cdot), \tilde{x}_{\mathcal{I}}(\cdot), \tilde{a}_{\mathcal{I}}(\cdot), \tilde{d}_{\mathcal{E}}(\cdot), \tilde{\xi}_{\mathcal{E}}(\cdot))$$

a *hydrodynamic model* of the system with initial state $(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0))$; a set $(\tilde{\psi}_{\mathcal{E}}(\cdot), \tilde{q}_{\mathcal{I}}(\cdot))$, which is a projection of a hydrodynamic model we often call a hydrodynamic model as well.

PROOF OF THEOREM 2.72. The non-trivial part of the proof is establishing that the limit $(\tilde{\psi}_{\mathcal{E}}(\cdot), \tilde{q}_{\mathcal{I}}(\cdot))$ is Lipschitz, which here is not a simple consequence of the functional law of large numbers for the driving processes (as was the case for fluid and hydrodynamic limits). This is because the arrival and service rates in the system with index r are $O(r)$, while the space is scaled down by $h(r) = o(r)$. For the same reason, it is also not "automatic" that the limit queues $\tilde{q}_i(\cdot)$ stay at 0. This difficulty is resolved as follows.

Consider an arbitrary number $C_4 > \Big\|(\tilde{\psi}_C E(0))\Big\|$, and the random time

$$\tau(r) = \min\{t : \Big\|(\tilde{\psi}^r_{ij}(t))\Big\| \geq C_4\}. \tag{70}$$

Speaking informally (the formal statements are given below), the trajectory $\tilde{x}^r_{\mathcal{I}}(\cdot)$ must be "almost Lipschitz" in the interval $[0, \tau(r)]$, with Lipschitz constant $\eta = C_4 \sum_{(ij)\in\mathcal{E}} \mu_{ij}$, because the absolute difference between the arrival and departure rates (scaled down by $h(r)$) is bounded above by η in $[0, \tau(r)]$. A similar observation holds for each queue length trajectory $\tilde{q}^r_i(\cdot)$, as long as the corresponding queue is non-zero. We will show that $\tau(r)$ is bounded away from 0 for all large r.

Suppose not, and $\tau(r) \to 0$ along some subsequence. Denoting $\tilde{x}_i(0) = \sum_j \tilde{\psi}_{ij}(0)$, we have

$$\sup_{[0,\tau(r)]} \|\tilde{x}^r_{\mathcal{I}}(t) - \tilde{x}_{\mathcal{I}}(0)\| \to 0, \quad \sup_{[0,\tau(r)]} \|\tilde{q}^r_{\mathcal{I}}(t) - \tilde{q}_{\mathcal{I}}(0)\| \to 0. \tag{71}$$

We also must have

$$\sup_{[0,\tau(r)]} \left\|\tilde{\psi}^r_{\mathcal{E}}(t) - \tilde{\psi}_{\mathcal{E}}(0)\right\| \to 0; \tag{72}$$

if not, we would be able to construct a hydrodynamic model which violates the condition that after a finite time the vector of occupancies $\overline{\psi}_{\mathcal{E}}(t)$ is uniquely determined as $L'\overline{x}_{\mathcal{I}}(t)$. However, (72) contradicts the definition of $\tau(r)$. We conclude that the case $\tau(r) \to 0$ is impossible, i.e. there exists some $\epsilon_4 > 0$ such that $\liminf \tau(r) > \epsilon_4 > 0$.

If $\liminf \tau(r) > \epsilon_4 > 0$ along some subsequence, then it is easy to see that there exists a further subsequence along which

$$\tilde{x}^r_{\mathcal{I}}(\cdot) \to \tilde{x}_{\mathcal{I}}(\cdot), \quad \tilde{q}^r_{\mathcal{I}}(\cdot) \to \tilde{q}_{\mathcal{I}}(\cdot), \tag{73}$$

where the convergences are uniform in $[0, \epsilon_4]$, and each function $\tilde{x}_i(\cdot)$ and $\tilde{q}_i(\cdot)$ is Lipschitz with constant η in $[0, \epsilon_4]$.

Next, in addition to (73), we show that

$$\left\|(\tilde{\psi}^r_{\mathcal{E}}(t), \tilde{q}^r_{\mathcal{I}}(t)) - L(\tilde{\psi}^r_{\mathcal{E}}(t), \tilde{q}^r_{\mathcal{I}}(t))\right\| \to 0, \text{ in particular } \left\|\tilde{\psi}^r_{\mathcal{E}}(t) - L'\tilde{x}^r_{\mathcal{I}}(t)\right\| \to 0, \tag{74}$$

uniformly in $[0, \epsilon_4]$. Suppose not; then we would be able to construct a hydrodynamic model which would violate the condition that

$$(\overline{\psi}_{\mathcal{E}}(t), \overline{q}_{\mathcal{I}}(t)) = L(\overline{\psi}_{\mathcal{E}}(t), \overline{q}_{\mathcal{I}}(t))$$

must hold after a finite time.

In $[0, \epsilon_4]$ we also have

$$\tilde{x}_i(t) = \tilde{x}_i(0) - \tilde{d}_i(t), \ \forall i,$$

where the Lipschitz function $\tilde{d}_i(\cdot)$ is a limit (along a subsequence) of

$$\sum_j \int_0^t \mu_{ij} \tilde{\psi}^r_{ij}(s) ds.$$

The above properties lead to conditions (69) on the interval $[0, \epsilon_4]$. Namely, we formally define $(\tilde{\psi}_{\mathcal{E}}(\cdot)) = L'(\tilde{x}_{\mathcal{I}}(\cdot))$, obtain the convergence $(\tilde{\psi}^r_{\mathcal{E}}(\cdot)) \to (\tilde{\psi}_{\mathcal{E}}(\cdot))$ from (74), and then (69) follows.

Conditions (69) reduce to a system of linear ordinary differential equations for $\tilde{\psi}_C E(t)$. In particular, each local fluid model remains bounded in $[0, \infty)$. This allows us to conclude that by choosing a sufficiently large C_4, the corresponding ϵ_4 (which bounds from below the time $\tau(r)$ taken for the state to leave the ball of radius C_4) can be arbitrarily large.

The fact that each local fluid model converges to 0 is easily established, again by induction on activities. Since the solution of a linear ODE converges exponentially quickly whenever it converges at all, the bound (68) follows. □

We will actually need a generalised version of Theorem 2.72.

THEOREM 2.74. *Consider a sequence of deterministic realisations, such that the driving realisations satisfy* (65)–(66). *Assume that the initial states converge to a fixed vector* $(\tilde{\psi}^r_{\mathcal{E}}(0), \tilde{q}^r_{\mathcal{I}}(0)) \to (\tilde{\psi}_{\mathcal{E}}, \tilde{q}_{\mathcal{I}})$. *(We do* not *assume* $(\tilde{\psi}_{\mathcal{E}}, \tilde{q}_{\mathcal{I}}) = L(\tilde{\psi}_{\mathcal{E}}, \tilde{q}_{\mathcal{I}})$.*) Then, for any subsequence of* r *there exists a further subsequence along which*

$$(\tilde{\psi}^r_{\mathcal{E}}(\cdot), \tilde{q}^r_{\mathcal{I}}(\cdot)) \to (\tilde{\psi}_{\mathcal{E}}(\cdot), \tilde{q}_{\mathcal{I}}(\cdot)), \tag{75}$$

in $D[\eta, \infty)$ *for any* $\eta > 0$, *where* $(\tilde{\psi}_{\mathcal{E}}(\cdot), \tilde{q}_{\mathcal{I}}(\cdot))$ *is a local fluid model with initial state*

$$(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0)) = L(\tilde{\psi}_{\mathcal{E}}, \tilde{q}_i).$$

Moreover, these limit trajectories are such that, uniformly on all of them,

$$\left\|(\tilde{\psi}_{\mathcal{E}}(t), \tilde{q}_{\mathcal{I}}(t))\right\| \le \left\|(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0))\right\| c_1 e^{-c_2 t}, \ \forall t \ge 0, \tag{76}$$

where $c_1, c_2 > 0$ *are fixed constants.*

The proof is a slight generalisation of that of Theorem 2.72. The initial jump in the local fluid model from $(\tilde{\psi}_{\mathcal{E}}, \tilde{q}_{\mathcal{I}})$ to $(\tilde{\psi}_{\mathcal{E}}(0), \tilde{q}_{\mathcal{I}}(0))$ is proved by considering an interval $[0, T_5 h(r)]$ and the corresponding hydrodynamic scaled trajectories in $[0, T_5]$; T_5 is chosen large enough so that the hydrodynamic model reaches the state

$$(\overline{\psi}_{\mathcal{E}}(0), \overline{q}_{\mathcal{I}}(0)) = L(\overline{\psi}_{\mathcal{E}}, \overline{q}_{\mathcal{I}})$$

by time T_5.

COROLLARY 2.75. *There exists* $C > 0$ *such that the following holds. For any fixed* $T > 0$, $K > 0$, $\delta_2 > 0$ *and* $\epsilon_2 > 0$, *there exists a sufficiently small* $\delta_3 > 0$, *such that: uniformly on all* $\left\|\tilde{f}^r(0)\right\| \le K$ *and all sufficiently large* r, *conditions*

$$\max_i \sup_{[0,T]} |h(r)^{-1}(A^r_i(t) - \lambda_i r t)| \le \delta_3, \tag{77}$$

$$\max_{(ij)} \sup_{[0,T]} |h(r)^{-1}\big(D^r_i(t) - \mu_{ij} \int_o^t \Psi^r_{ij}(s) ds\big)| \le \delta_3, \tag{78}$$

imply

$$\sup_{[0,T]} \left\|\tilde{f}^r(t)\right\| \le (K+1)C, \tag{79}$$

$$\sup_{[\epsilon_2, T]} \left\|\tilde{f}^r(t) - \tilde{f}(t)\right\| \le \delta_2, \tag{80}$$

where $\tilde{f}(\cdot)$ *is a local fluid model with initial state* $L\tilde{f}^r(0)$ *(so that* $\tilde{f}(\cdot)$ *depends on* r*).*

8.4. Proof of Theorem 2.64(ii). We are now in position to prove (56), and then Theorem 2.64(ii). The basic idea is to consider the process in the interval $[0, T \log r]$, subdivided into $\log r$ intervals[15], each being T-long. Using the local fluid limit results, we show that, with high probability, in each of the T-long subintervals, the norm $\|F^r(t)\|$ decreases by a factor $\delta_6 \in (0,1)$, unless the norm $\|F^r(t)\|$ at the beginning of the subinterval was smaller than $r^{1/2+\epsilon}$; in this case $\|F^r(t)\|$ will be bounded above by $3Cr^{1/2+\epsilon}$ during the entire subinterval (where C is as in Corollary 2.75). If δ_6 is small enough, so that

$$\delta_6^{\log r} < r^{1/2+\epsilon}/r, \quad \delta_6 < e^{-1/2+\epsilon}, \tag{81}$$

then the above implies $\|F^r(t)\|$ must "dip" below $r^{1/2+\epsilon}$ at least once, and therefore $\|F^r(T \log r)\| \le 3Cr^{1/2+\epsilon}$ (with high probability). We proceed with the details.

[15]To be precise, we should consider an integer number of subintervals, say $\lfloor \log r \rfloor$. This does not cause any difficulties besides making notation cumbersome.

Let us choose $\delta_6 > 0$ satisfying (81), and then $\delta_2 > 0$ such that $2\delta_2 < \delta_6$. Denote by $\|L\|$ the norm of the linear operator L (defined in Theorem 2.71), i.e. the maximum of the absolute values of its eigenvalues. Let us choose $T > 0$ large enough so that (see Theorem 2.74) $\|L\| c_1 e^{-c_2 T} < \delta_2$.

Suppose, for each r the initial state is as in (56). To prove (56) it suffices to show that from any subsequence of r we can find a further subsequence, along which (56) holds. So, consider any fixed subsequence, and a fixed $\delta_1 > 0$.

In each of the subintervals $[(i-1)T, iT]$, $i = 1, 2, \ldots, \log r$, we consider the process with the time origin reset to $(i-1)T$ and the corresponding initial state $F^r((i-1)T)$. If $\|F^r((i-1)T)\| \leq g(r)$, then we set $h(r) = \max(\|F^r((i-1)T)\|, r^{1/2+\epsilon})$; if $\|F^r((i-1)T)\| > g(r)$ we set $h(r) = g(r)$ for completeness, but with high probability this will never occur. By Proposition 2.66, we can choose a further subsequence so that, w.p.1, conditions (77) and (78) hold for all large r, *simultaneously* on each of the subintervals $[0, T]$, $[T, 2T]$, $\ldots$, $[T(\log r - 1), T \log r]$. We consider the corresponding local fluid scaled processes $\tilde{f}^r(\cdot)$, with their corresponding $h(r)$, on each of the subintervals; and apply Corollary 2.75. We see that, with probability 1, for all large r, the following holds for each interval $[(i-1)T, iT]$, $i = 1, 2, \ldots, \log r$:
if $\|F^r((i-1)T)\| \in [r^{1/2+\epsilon}, g(r)]$ then $\|F^r(iT)\| \leq 2\delta_2 \|F^r((i-1)T)\|$;
if $\|F^r((i-1)T)\| < r^{1/2+\epsilon}$ then $\|F^r(iT)\| \leq 3Cr^{1/2+\epsilon}$.

Since $2\delta_2 < \delta_6$ we must have $\|F^r(iT)\| < r^{1/2+\epsilon}$ for at least one i. Finally, we conclude that the condition $\|F^r(T \log r)\| \leq 3Cr^{1/2+\epsilon}$ must hold (w.p.1 for all large r). This obviously implies (56).

We believe that a stronger result is also true.

Conjecture 2.76. The sequence of stationary distributions of the processes

$$r^{-1/2}(\Psi^r_{\mathcal{E}}(\cdot) - r\psi^*_{\mathcal{E}}, Q^r_{\mathcal{I}}(\cdot))$$

is tight.

CHAPTER 3

Limit order book

Introduction

In this chapter we model a limit order book. A limit order book is a pricing mechanism for a single-commodity market. To illustrate the concept of a pricing mechanism, suppose you would like to buy a carrot. Depending on the amount of time and money you have (and the amount of ridicule you're willing to put up with), there are several ways in which you could go about acquiring it:

- Go to a supermarket, and pay the price of a carrot written on the shelf.
- Go to a farm, and haggle over the price with the farmer.
- Go to a street market and in a booming voice announce "I need a carrot; who will offer me the best price?" in the hopes that this will spur the stall-keepers into a price war.
- Bid on a carrot on eBay.
- ...

All of these are mechanisms for pairing up buyers and sellers (of carrots), and for deciding the amount of money that will be exchanged during the transaction.

In these terms, a limit order book works as follows. Sellers and buyers of carrots arrive in real time. They publicly make one of the following four announcements:

- I would like to sell a carrot right now, to the highest waiting bidder in the system. ("Market ask")
- I would like to buy a carrot right now, from the lowest seller in the system. ("Market bid")
- I have some carrots, and could be persuaded to part with them, but only if the price rises above p. ("Limit ask")
- I would like to invest in some carrots eventually, but only if the price drops below p. ("Limit bid")

The market bid and market ask are essentially equivalent to trading with a supermarket: you get to buy or sell the carrot immediately, but possibly at an inconveniently high, respectively low, price. The limit orders, on the other hand, may result in better deals, but involve waiting for the order to be executed. The limit order book is the list of unfulfilled limit bids and limit asks.

Although somewhat impractical as a way of getting a single carrot for dinner, this pricing mechanism is important in financial markets, many of which are run using variants of this model. Consequently, it has generated a lot of interesting research, both empirical (studies of real-world market data) and theoretical (models of how the behaviour might arise). The following discussion is taken from the excellent survey of [Gould et al., 2011], and many more references can be found therein.

Empirical studies. Because the information (prices and sizes of orders) in the limit order book is publicly available[1], quite a lot of statistical data about limit order book

[1]With caveats: not all of the limit order book is available to the general public, there may be a delay, some of the markets allow asymmetric information, and in some markets it is possible to submit partially or completely hidden orders. The hidden orders in particular greatly complicate empirical studies.

behaviour has been amassed. Many of the empirical studies contradict each other, possibly due to fundamental differences between the underlying markets, or due to the inherent difficulty of the problem. However, there are a few features shared by many markets. The time series of prices[2] has certain interesting characteristics, including different *volatilities* at different time scales. There are several ways to define volatility, but loosely speaking, volatility is a measure of variability of the logarithm of the price over that time scale. For example, to compute the 5-minute volatility, one could look at the series of prices $p(t_0), p(t_1), \ldots$ spaced by 5 minutes, and compute the standard deviation of the quantity $\log p(t_{i+1}) - \log p(t_i)$. One can also consider the time series of volatilities on a given time scale, e.g., the day-by-day 5-minute volatility. Observations suggest that high-volatility periods tend to cluster together, as do low-volatility periods; that is, large variations in prices are more likely to follow other large variations in prices than they are to occur unconditionally.

Another feature found in many markets is the "humped" shape of the limit order book. Here, we consider the total quantity of the good being offered for sale, or requested, as a function of price. Many studies find that each of the buy and sell distributions are approximately unimodal, with the maximum occurring at some price that is near, but not equal to, the current best price. Finally, some studies find that the process describing the limit order book may not be stationary; this is interpreted as the result of the new information being constantly supplied to the market. It may also be possible to model this as evolving "steady-state behaviour" of a system whose underlying parameters vary over time.

The theoretical models of limit order books have largely fallen into the following two classes.

Economic game theory. A limit order book can be naturally modelled as a large repeated game, in which players have more or less information about each others' preferences. Two studies using this set-up are [Parlour, 1998] and [Roşu, 2009]. In [Parlour, 1998], orders cannot be changed after placement, and the set of possible prices is reduced to just two ticks ("high" and "low", corresponding to current selling price and buying price). Thus, the strategic choices are essentially "place a market order", "join the queue of limit orders", or "pass". This models the trade-off between the price and the probability of an order being executed before a deadline. Roşu [2009] introduces the possibility of modifying orders after they are submitted. This and the assumption of continuously-varying prices turns out to simplify the space of possible strategies enough to derive the form of the subgame-perfect Nash equilibria for the system. Both models assume a large amount of (symmetric) common knowledge available to all the market participants; for example, everyone knows everyone else's level of aversion to waiting. The strategies giving the game-theoretic equilibria of these two models explain some of the features of real-world limit order book markets. While it is possible that a fuller model of this flavour would explain more of the behaviour, analysing a large repeated game in continuous time is tricky at best.

Zero-knowledge and Markovian markets. Because modelling individual buyers is difficult, one could try to model the market without referring to the individual buyer and seller preferences, and instead specifying stochastic dynamics for the market as a whole. An early paper introducing these ideas is [Gode and Sunder, 1993]; they consider a small market with *zero-intelligence* traders is considered. Zero-intelligence traders make their decisions based only on the current price, without attempting to game the system

[2]There isn't a single well-defined *price* in the limit order book; so this could refer to the *highest bid* price, the *lowest ask* price, the *mid-point* price halfway between the two, or the price at which the most recent transaction occurred.

in any way. One interesting feature that emerges is a notion of an equilibrium price: even through the traders do not "discuss" their different valuations in any way, trades in the market eventually only occur around some single price. An example of a Markovian market is given in [Cont and de Larrard, 2010], which uses a Markov process to represent the market state. Cont and de Larrard [2010] assume that the Markovian state descriptor can be taken to be just the pair (bid price, ask price), rather than the full state of the limit order book. Within this model, the authors are able to derive steady-state distributions of various quantities, such as price movements.

A trend in literature is to add assumptions to the model until it reproduces the desired statistical properties of the real-world limit order books. 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 numerical analysis, which makes it difficult to understand the effect that the parameters of the model have on its behaviour.

The analysis in this chapter is instead a deliberately very simple and almost parameter-free model. This is because we do not set the goal of approximating "real-world behaviour" as closely as possible. Rather, we would like to understand the behaviour of the underlying system of interacting queues, in the hopes that the insights will generalise to other settings. Consequently, there are very few adjustable parameters in the model we analyse, although we show some possible extensions in §10 and §9. It is interesting that even in such a simple system nontrivial behaviour emerges; for example, we see clear threshold values for orders clearing from the system.

1. Limit order book model

Definition 3.1. An *order* is a pair (price, type). Price is a real number; type is one of "bid" and "ask". A *bid* is an order to buy a unit of good (one carrot, in the terminology of the introduction); an *ask* is an order to sell one unit of good. Orders arrive exogenously into the limit order book, and cannot be cancelled: they are either *executed*, that is matched to an order of the opposite type (immediately or at a later time), or they remain in the system forever.

We are implicitly assuming that all orders have the same size; this is a simplification (in real life, you may want to buy not one but a dozen carrots). Roşu [2009] discusses the effect that the order sizes may have on the statistical properties of real-world limit order books. The assumption that orders cannot be cancelled is also a simplification: in real life, many if not most of the orders are indeed cancelled before execution (and perhaps submitted without intending for them to be executed). We can think of our model as applying to the orders that really are meant to be executed; but the primary reason for the assumption is that the monotonicity results in Section 3 break if we allow orders to depart at will.

Definition 3.2. The state of a limit order book at time t is a pair of counting measures $(\mathfrak{Q}_t^b, \mathfrak{Q}_t^a)$ supported on $[0,1]$, counting bids and asks respectively, with cumulative density functions $Q_t^b(p) \equiv \mathfrak{Q}_t^b(-\infty, p]$ and $Q_t^a(p) \equiv \mathfrak{Q}_t^a[p, \infty)$ (note that asks are counted from the right).

The quantities $\mathfrak{Q}_t^b\{p\}$ and $\mathfrak{Q}_t^a\{p\}$ are known in the financial literature as the *depth* of the market at price p; when p is, e.g., the highest bid price, this is the number of transactions that need to occur in order to change the price.

We will occasionally want to consider infinite states, but only "nice" ones. Specifically, all states we will consider have finite support, i.e. for $i = a, b$, there are only finitely many prices p such that $\mathfrak{Q}_t^i\{p\} \neq 0$. However, we may have $\mathfrak{Q}^b\{p\} = \infty$ and/or $\mathfrak{Q}^a\{q\} = \infty$

for some prices p and q. If we have both infinitely many bids at p and infinitely many asks at q, we will always have $p < q$,[3] so that we never have to consider infinitely many departures in a finite time period.

The external arrivals to the limit order book happen according to some processes $(A_t)_{t\geq 0}$; the set of times at which an arrival occurs is discrete. Each arrival event is an order, i.e. a pair (price, type). Unless otherwise noted, arrivals happen in discrete time, are iid, the type is equally likely to be "bid" or "ask", and the price is supported on $[0, 1]$. (We may also consider Poisson arrival processes, in which case we will restrict to the probability-1 event that there are only finitely many arrivals in any finite time interval, and no two arrival times coincide.) For most of this chapter, we will assume that the distribution of the price of an arriving order is the same for bids and for asks, in which case it can without loss of generality be taken to be uniform on $[0, 1]$ (see §9, where we also discuss what happens if the arrivals are iid but with different distributions for bids and asks).

The departures from an order book happen only at arrival times. Informally, an arriving bid departs if there is an ask in the system to the left of it, and in that case it departs with the leftmost such ask; similarly, an arriving ask departs if there is a bid in the system to the right of it, and in that case it departs with the rightmost such bid. We will also consider the effect of partitioning prices into discrete ticks, which we will formalise by introducing price level functions.

Definition 3.3. A *price level function*, which we may also refer to as a *pricing scheme*, called $\mathcal{P}$ and denoted $x \prec y$, is a partial ordering on $[0, 1]$ that is refined by the usual total ordering. Equivalently, it is a nondecreasing map $\mathcal{P} : [0, 1] \to [0, 1]$, where we define $p \prec q$ to mean $\mathcal{P}(p) < \mathcal{P}(q)$. (We will always take this map to be right-continuous.) When p is incomparable to q (i.e., none of $p \prec q$, $p = q$, or $p \succ q$ hold), we write $p \sim q$.

We allow the highest bid-lowest ask pair to depart the system whenever $\beta_t \not\prec \alpha_t$ (note that one of β_t and α_t in this case must be a newly arrived order).

To formally specify the dynamics of the system, we will introduce two more pairs counting measures, counting the cumulative arrivals and the cumulative departures from a set.

Definition 3.4. For a set of prices S, let $\mathfrak{A}^a_t(S)$ ($\mathfrak{A}^b_t(S)$) denote the number of asks (bids) with prices in S that have arrived into the system by time t, and let $\mathfrak{D}^a_t(S)$ ($\mathfrak{D}^b_t(S)$) denote the number of asks (bids) with prices in S that have left the system up to time t. We will always consider starting states for which $\mathfrak{D}^a_0 = \mathfrak{D}^b_0 = 0$ is the zero measure. Let $D^b_t(p) = \mathfrak{D}^b_t(-\infty, p]$ and $D^a_t(p) = \mathfrak{D}^a_t[p, \infty)$ be the cumulative distribution functions for the departure measures; note that asks are counted from the right.

We formally define the evolution of a limit order book with price level function $\mathcal{P}$: Upon the arrival at time t of a bid at price p at time t, if $\alpha_{t^-} \succ p$, then the bid waits:

$$\mathfrak{A}^b_t = \mathfrak{A}^b_{t^-} + \delta_p, \quad \mathfrak{A}^a_t = \mathfrak{A}^a_{t^-}, \quad \text{if } A_t = (p, \text{bid}) \text{ and } \alpha_{t^-} \succ p. \tag{82a}$$

If $\alpha_{t^-} \not\succ p$, then the bid departs with the leftmost ask:

$$\mathfrak{A}^a_t = \mathfrak{A}^a_{t^-} - \delta_{\alpha_{t^-}}, \quad \mathfrak{A}^b_t = \mathfrak{A}^b_{t^-}, \quad \text{if } A_t = (p, \text{bid}) \text{ and } \alpha_{t^-} \not\succ p.$$

$$\mathfrak{D}^b_t = \mathfrak{D}^b_{t^-} + \delta_p, \quad \mathfrak{D}^a_t = \mathfrak{D}^a_{t^-} + \delta_{\alpha_{t^-}} \tag{82b}$$

The situation is symmetrical if the order arriving at time t is an ask at price p: if $\beta_{t^-} \prec p$ then the ask waits, while if $\beta_{t^-} \not\prec p$, then both the ask and the rightmost bid depart.

[3] More precisely, $p \prec q$ in the appropriate partial ordering $\prec$; see below.

The relationship between the quantities $\mathfrak{Q}$, $\mathfrak{A}$, and $\mathfrak{D}$ is as follows: for a set of prices S, any time $t \geq 0$, and $i = a, b$,

$$\mathfrak{Q}_t^i(S) = \mathfrak{Q}_0^i(S) + \mathfrak{A}_t^i(S) - \mathfrak{D}_t^i(S). \tag{83}$$

We require that $\beta_0 \prec \alpha_0$, so that no departures are possible initially. The evolution described in (82) guarantees that all of these quantities are finite, and $\beta_t \prec \alpha_t$ at all times $t \geq 0$.

2. Main results

In this section we state our main results. Their proofs will be given in later sections: the proof of Theorems 3.5 and 3.7 is in §4; the proof of Theorems 3.9 and 3.11 is in §6; the proof of Theorem 3.10 is in §8.

THEOREM 3.5. *Let L be a limit order book with deterministic starting state $(\mathfrak{Q}_0^b, \mathfrak{Q}_0^a)$ and arrival process $(A_t)_{t\geq 0}$. Suppose that the arrival events are independent. Then there exist two constants κ_b and κ_a such that the following hold for any $\epsilon > 0$, with probability 1.*

1. *$\mathfrak{D}_\infty^b(-\infty, \kappa_b - \epsilon] < \infty$, and $\mathfrak{D}_\infty^a[\kappa_a + \epsilon, \infty) < \infty$. That is, only finitely many bid departures at prices $< \kappa_b - \epsilon$ ever occur, and only finitely many ask departures at prices $> \kappa_a + \epsilon$ ever occur.*
2. *The event $\{\mathfrak{Q}_t^b[\kappa_b+\epsilon, \infty) = 0\}$ occurs infinitely often, and the event $\{\mathfrak{Q}_t^a(-\infty, \kappa_a - \epsilon] = 0\}$ occurs infinitely often. That is, infinitely often all of the bids at prices $> \kappa_b + \epsilon$ are executed, and infinitely often all of the bids at prices $< \kappa_a - \epsilon$ are executed.*

Further, the constants κ_b and κ_a do not change if the starting state $(\mathfrak{Q}_0^b, \mathfrak{Q}_0^a)$ is modified by a finite number of bids.

DEFINITION 3.6. We call κ_b and κ_a in Theorem 3.5 the *threshold values* on the bid and ask side respectively.

When the arrivals are iid with some bounded density, we have the following refinement:

THEOREM 3.7. *Let L be a limit order book with some deterministic finite starting state, arbitrary price level function, and arrival process $(A_t)_{t\geq 0}$. Let the arrival events $(A_t)_{t\geq 0}$ be iid, with*

$$\mathbb{P}(A_t \in dp \times bid) = \frac{1}{2} dF^b(p), \quad \mathbb{P}(A_t \in dp \times ask) = \frac{1}{2} dF^a(p)$$

for some pair of probability distributions F^b, F^a on $[0, 1]$ with bounded densities f^b, f^a respectively; let

$$M = \max_{i=a,b} \sup_{p\in[0,1]} f^i(p)$$

Then the threshold values κ_b and κ_a satisfy $F^b(\kappa_b) = 1 - F^a(\kappa_a)$. Moreover, for any $\epsilon > 0$, w.p.1, there exists a sequence of times $T_n \to \infty$ such that $\mathfrak{Q}_{T_n}^b[\kappa_b + \epsilon, \infty) = 0$ and

$$\limsup_{T_n\to\infty} \frac{1}{T_n} \mathfrak{Q}_{T_n}^a(-\infty, \kappa_a + \epsilon] \leq 2M\epsilon.$$

REMARK 3.8. The boundedness of the densities is used to control the number of bid or ask arrivals on a small interval near the boundary values κ_b, κ_a. In particular, we only require the density to be bounded on a neighbourhood of κ_b and κ_a.

In the above two theorems, the event of measure 1 on which the results hold may depend on ϵ. In Theorem 3.7, the sequence of times T_n is random as well: that is, for almost every ω in the underlying probability space there exists a sequence $T_n = T_n(\omega)$ satisfying the conditions of the theorem.

Theorem 3.5 is applicable to a very wide class of arrival processes; and it leaves open the possibility that $\kappa_b = 0$ and $\kappa_a = 1$, or that $\kappa_b = \kappa_a$. Theorem 3.9 shows that, when the arrivals are iid uniform (and the partial ordering is not too coarse), this is not the case: we really do have a positive fraction of unfulfilled bid and ask orders. Theorem 3.10 shows that $\kappa_b < \kappa_a$, so there is a nontrivial region where all bids periodically clear, and all asks periodically clear. (We do not know whether all orders will clear infinitely often on the entirety of this region.) Finally, Theorem 3.11 computes the threshold values precisely for the case of uniform arrivals and continuous pricing.

THEOREM 3.9. *Let L be a limit order book with some deterministic finite starting state, price level function $\mathcal{P}$, and arrival process $(A_t)_{t\geq 0}$. Suppose the arrival events be iid uniform on $[0,1] \times \{bid, ask\}$, and suppose $\mathcal{P}$ is such that there exists a price p with $0 \prec p \prec 1-p \prec 1$. Then*

$$\frac{p-2p^2}{2-3p} \leq \kappa_b, \quad \frac{p-2p^2}{2-3p} \leq 1-\kappa_a.$$

In particular, $\frac{1}{9} \leq \kappa_b$ and $\kappa_a \leq \frac{8}{9}$ when $\mathcal{P}(x) = x$.

THEOREM 3.10. *Let L be a limit order book with some deterministic finite starting state, arbitrary price level function $\mathcal{P}$, and arrival process $(A_t)_{t\geq 0}$. Suppose the arrival events be iid uniform on $[0,1] \times \{bid, ask\}$. Then $\kappa_b \leq \frac{1}{4}$ and $1-\kappa_a \leq \frac{1}{4}$.*

For the case of $\mathcal{P}(x) = x$, we can find the value of κ_b and κ_a precisely.

THEOREM 3.11. *Let the partial ordering $\mathcal{P}$ be $\leq$, i.e. given by $\mathcal{P}(x) = x$. Let the arrival events be iid uniform on $[0,1] \times \{bid, ask\}$. Then the value of κ_b is given as the unique solution to*

$$\log\left(\frac{1-\kappa_b}{\kappa_b}\right) = \frac{\kappa_b}{1-\kappa_b} + 1, \quad \kappa_b \approx 0.2178$$

and $\kappa_a = 1 - \kappa_b$.

3. Monotonicity

In this section we gather some of the monotonicity results that our model exhibits. First, we examine what happens if we modify the starting state of the model, i.e. what effect will a change in the initial configuration have on the future evolution of the limit order book. Second, we relax the pricing scheme; i.e. we replace the price level function $\mathcal{P}$ by $\tilde{\mathcal{P}}$, where $x \tilde{\prec} y$ implies $x \prec y$.

REMARK 3.12. There is a body of work on proving monotonicity for Markov processes; see for example [Massey, 1987], or more recently [Delgado et al., 2004] and [Lorek and Szekli, 2012], and references therein. In our case, the proofs are sufficiently simple to simply derive from scratch.

LEMMA 3.13. *Let L and $\tilde{L}$ be two limit order books sharing the same arrival process and price level function, but such that*

$$\tilde{\mathfrak{Q}}_0^b = \mathfrak{Q}_0^b + \delta_{p_0} \text{ for some price } p_0, \quad \tilde{\mathfrak{Q}}_0^a = \mathfrak{Q}_0^a.$$

Then at all subsequent times either $\tilde{\mathfrak{Q}}_t^b = \mathfrak{Q}_t^b + \delta_{p_t}$ and $\tilde{\mathfrak{Q}}_t^a = \mathfrak{Q}_t^a$, or $\tilde{\mathfrak{Q}}_t^b = \mathfrak{Q}_t^b$ and $\tilde{\mathfrak{Q}}_t^a = \mathfrak{Q}_t^a - \delta_{q_t}$, for some prices p_t, q_t (which depend on t, and may be random).

PROOF. Up until the time the extra bid has departed, the departures are exactly the same in the two systems, so there is an extra bid in $\tilde{L}$. When the extra bid does depart, there is one fewer ask in $\tilde{L}$. We now repeat the argument swapping the roles of L and $\tilde{L}$, and of bids and asks. □

COROLLARY 3.14. *Let $\tilde{L}$ be obtained from L by adding some bids at some set of times; except for that change, the starting state, arrival process, and price level function of L and $\tilde{L}$ coincide. Then at all times, $\tilde{L}$ has at least as many bids as L and no more asks than L.*

COROLLARY 3.15. *Let $\tilde{L}$ be obtained from L by modifying the starting state by a finite number $\leq M$ of orders; except for that change, the starting state, arrival process, and price level function of L and $\tilde{L}$ coincide. Then the states of L and $\tilde{L}$ at all times will differ by at most M orders. In particular, for any set S and $i = a, b$ we will have $\left|\mathfrak{D}_t^i(S) - \tilde{\mathfrak{D}}_t^i(S)\right| \leq M$.*

We next discuss different pricing schemes.

DEFINITION 3.16. For two price level functions $\mathcal{P}$, $\tilde{\mathcal{P}}$ we say that $\tilde{\mathcal{P}}$ is *coarser* than $\mathcal{P}$ if $x \tilde{\prec} y$ implies $x \prec y$.

Relaxing the pricing scheme into a coarser one means that more bid-ask pairs become "eligible" to leave; Lemma 3.17 asserts that more pairs really do leave.

LEMMA 3.17. *Let L and $\tilde{L}$ be limit order books with the same starting state and external arrival process, but let $\tilde{\mathcal{P}}$ be coarser than $\mathcal{P}$. Then $\tilde{D}_t^b(p) \geq D_t^b(p)$ and $\tilde{D}_t^a(p) \geq D_t^a(p)$ for all prices p and times t.*

We will need a preliminary result, which is essentially an observation about increasing functions.

LEMMA 3.18. *Let L, $\tilde{L}$ be limit order books as in Lemma 3.17, and suppose that at some time t we have $\tilde{D}_t^b(p) \geq D_t^b(p)$ and $\tilde{D}_t^a(p) \geq D_t^a(p)$ for all p. Further, suppose that $\tilde{D}_t^b(\infty) = D_t^b(\infty)$. Then the rightmost bid satisfies $\tilde{\beta}_t \geq \beta_t$, and the leftmost ask satisfies $\tilde{\alpha}_t \leq \alpha_t$.*

PROOF. First, observe that $\tilde{D}_t^b(\infty) = \tilde{D}_t^a(-\infty)$ and $D_t^b(\infty) = D_t^a(-\infty)$, since bids and asks depart in pairs. We prove the statement about the rightmost bid; the statement about the leftmost ask will follow by an identical argument.

Suppose that the lemma does not hold, i.e. the set $[\beta_t, \infty)$ has $\mathfrak{Q}_t^b[\beta_t, \infty) > 0$ but $\tilde{\mathfrak{Q}}_t^b[\beta_t, \infty) = 0$ (i.e., $\tilde{\beta}_t < \beta_t$). Since bids are only present at a discrete set of prices, let $\epsilon > 0$ be such that $\mathfrak{Q}_t^b(\beta_t - \epsilon, \infty) = \mathfrak{Q}_t^b[\beta_t, \infty)$ and similarly for $\tilde{\mathfrak{Q}}_t^b$.

From (83) we infer that

$$\tilde{\mathfrak{D}}_t^b(\beta_t - \epsilon, \infty) > \mathfrak{D}_t^b(\beta_t - \epsilon, \infty),$$

since the arrivals are equal. However, then

$$\tilde{D}_t^b(\beta_t - \epsilon) = \tilde{D}_t^b(\infty) - \tilde{\mathfrak{D}}_t^b(\beta_t - \epsilon, \infty) < D_t^b(\beta_t - \epsilon),$$

contradicting the assumptions of the lemma. □

PROOF OF LEMMA 3.17. The inequalities clearly hold at $t = 0$, and then we only need to check that they are preserved after an arrival. We will prove only the inequality for D^b; the proof for D^a is identical.

Bid arrival. We consider first the arrival event $A_t = (q, \text{bid})$. If $\tilde{\alpha}_{t^-} \succ q$ (bid doesn't depart in L), or if $\tilde{D}_{t^-}^b(p) > D_{t^-}^b(p)$ (strictly) for all $p \geq q$, then the (nonstrict) inequality

$\tilde{D}^b_t(p) \geq D^b_t(p)$ holds. Thus, we need to consider the case $\alpha_{t^-} \not\succ q$ and $\tilde{D}^b_{t^-}(p) = D^b_{t^-}(p)$ for some $p \geq q$.

Since there is an ask at $\alpha_{t^-} \not\succ q \leq p$ in L, there must be no bids to the right of p: $\mathfrak{Q}^b_{t^-}[p, \infty) = 0$. By (83) this implies that

$$\mathfrak{D}^b_{t^-}[p, \infty) = \mathfrak{Q}^b_0[p, \infty) + \mathfrak{A}^b_{t^-}[p, \infty)$$

has the biggest value it could possibly have. Since $\tilde{D}^b_{t^-}(p) = D^b_{t^-}(p)$, we see

$$D^b_{t^-}(\infty) = D^b_{t^-}(p) + \mathfrak{D}^b_{t^-}[p, \infty) \geq \tilde{D}^b_{t^-}(\tilde{D}^b_{t^-}(\infty).$$

Since we had assumed $D^b_{t^-}(\infty) \leq \tilde{D}^b_{t^-}(\infty)$, we must in fact have equality: that is,

$$\tilde{D}^b_{t^-}(\infty) = D^b_{t^-}(\infty).$$

Applying Lemma 3.18, we conclude $\tilde{\alpha}_{t^-} \leq \alpha_{t^-} \not\succ q$, and hence $\tilde{\alpha}_{t^-} \tilde{\not\succ} q$. Therefore, the arriving bid departs in both systems ($\mathfrak{D}^b_t = \mathfrak{D}^b_{t^-} + \delta_q$ and $\tilde{\mathfrak{D}}^b_t = \tilde{\mathfrak{D}}^b_{t^-} + \delta_q$) and the inequality is preserved.

Ask arrival. We now consider the arrival of an ask at price q. The argument will be very similar, except at the very end. If $\beta_{t^-} \prec q$ (no bid departs in L), or if $\tilde{D}^b_{t^-}(p) > \tilde{D}^b_{t^-}(p)$ (strictly) for all $p \geq \beta_{t^-}$, then the (nonstrict) inequality $\tilde{D}^b_t(p) \geq \tilde{D}^b_t(p)$ holds. Thus, we need to consider the case $\beta_{t^-} \not\prec q$ and $\tilde{D}^b_{t^-}(p) = D^b_{t^-}(p)$ for some $p \geq \beta_{t^-} \not\prec q$.

Since the rightmost bid in L is at β_{t^-}, we have $\mathfrak{Q}^b_{t^-}(\beta_{t^-}, \infty) = 0$, and in particular $\mathfrak{Q}^b_{t^-}[p, \infty) = 0$. By (83) this implies that

$$\mathfrak{D}^b_{t^-}[p, \infty) = \mathfrak{Q}^b_0[p, \infty) + \mathfrak{A}^b_{t^-}[p, \infty)$$

has the biggest value it could possibly have; since $\tilde{D}^b_{t^-}(p) = D^b_{t^-}(p)$, we see that

$$D^b_{t^-}(\infty) = D^b_{t^-}(p) + \mathfrak{D}^b_{t^-}[p, \infty) \geq \tilde{D}^b_{t^-}(\infty).$$

Since we had assumed $D^b_{t^-}(\infty) \leq \tilde{D}^b_{t^-}(\infty)$, we must in fact have equality: that is,

$$\tilde{D}^b_{t^-}(\infty) = D^b_{t^-}(\infty).$$

Applying Lemma 3.18, we conclude $\tilde{\beta}_{t^-} \geq \beta_{t^-} \not\prec q$, hence $\tilde{\beta}_{t^-} \tilde{\not\prec} q$. Thus, the arriving ask will depart in both systems (with a bid at $\tilde{\beta}_t$ in L, and with a bid at $\tilde{\beta}_{t^-}$ in $\tilde{L}$). This immediately implies

$$\tilde{D}^b_t(x) \geq D^b_t(x), \quad x \notin [\beta_{t^-}, \tilde{\beta}_{t^-}). \tag{84}$$

We claim that on the interval $[\beta_{t^-}, \tilde{\beta}_{t^-})$ we had a strict inequality

$$\tilde{D}^b_{t^-}(x) > D^b_{t^-}(x), \quad \text{for } x \in [\beta_{t^-}, \tilde{\beta}_{t^-}).$$

Indeed, for $x \in [\beta_{t^-}, \tilde{\beta}_{t^-})$, (83) implies

$$\mathfrak{D}^b_{t^-}(x, \infty) = \mathfrak{Q}^b_0(x, \infty) + \mathfrak{A}^b_{t^-}(x, \infty), \quad x \in [\beta_{t^-}, \tilde{\beta}_{t^-})$$

which is the biggest value it could possibly have. On the other hand,

$$\tilde{\mathfrak{D}}^b_{t^-}(x, \infty) = \mathfrak{Q}^b_0(x, \infty) + \mathfrak{A}^b_{t^-}(x, \infty) - \tilde{\mathfrak{Q}}^b_{t^-}(x, \infty) \leq \mathfrak{D}^b_{t^-}(x, \infty) - 1$$

on $[\beta_{t^-}, \tilde{\beta}_{t^-})$. The last inequality holds because $\tilde{\mathfrak{Q}}^b_{t^-}(x, \infty) \geq 1$, since by assumption there is a waiting bid at $\tilde{\beta}_{t^-} \geq x$ in $\tilde{L}$. Since we already know $D^b_{t^-}(\infty) = \tilde{D}^b_{t^-}(\infty)$, this inequality is enough to conclude

$$\tilde{D}^b_{t^-}(x) = \tilde{D}^b_{t^-}(\infty) - \tilde{\mathfrak{D}}^b_{t^-}(x, \infty) > D^b_{t^-}(x), \quad x \in [\beta_{t^-}, \tilde{\beta}_{t^-}). \tag{85}$$

Combining (84) and (85) yields the result. □

COROLLARY 3.19. *Let L and $\tilde{L}$ be limit order books with the same starting state and external arrival process, but let $\tilde{\mathcal{P}}$ be coarser than $\mathcal{P}$ (Definition 3.16). Then $\tilde{\kappa}_b \le \kappa_b$ and $\tilde{\kappa}_a \ge \kappa_a$.*

4. Proof of Theorems 3.5 and 3.7

In this section we prove Theorem 3.5. We will be using the machinery of the Kolmogorov 0-1 law for tail σ-algebras Williams [1991].

Define the events

$$\mathcal{A}^b(x) \equiv \{\mathfrak{D}^b_\infty(-\infty, x] < \infty\}, \quad \mathcal{A}^a(x) \equiv \{\mathfrak{D}^a_\infty[x, \infty) < \infty\}.$$

Note that for any set S the functions $\mathfrak{D}^i_t(S)$, $i = a, b$, are nondecreasing, so the limits as $t \to \infty$ always exist (but may be infinite).

For limit order books $\tilde{L}$, $\hat{L}$ we will denote the corresponding events $\tilde{\mathcal{A}}^i$, $\hat{\mathcal{A}}^i$ $(i = a, b)$.

PROOF OF THEOREM 3.5. Without loss of generality, we may index the time by nonnegative integers. Let $\mathcal{F}_n = \sigma(\{A_t, t \ge n\})$; the tail σ-algebra is $\mathcal{F} \equiv \bigcap_n \mathcal{F}_n$. We begin by showing that for any x, the events $\mathcal{A}^b(x)$, $\mathcal{A}^a(x)$ are $\mathcal{F}$-measurable, that is, that they are $\mathcal{F}_n$-measurable for all n. Below we consider $\mathcal{A}^b(x)$; the case of $\mathcal{A}^a(x)$ is similar.

The event $\mathcal{A}^b(x) \in \mathcal{F}_0$ because

$$\mathcal{A}^b(x) = \bigcup_m \bigcap_n \mathcal{A}^b_{n,m}(x)$$

where $\mathcal{A}^b_{n,m}(x)$ is the event that there are at most m bid departures at prices $p < x$ by the time of the n^{th} arrival (clearly, an element of $\mathcal{F}_0$).

We now show that $\mathcal{A}^b(x)$ is $\mathcal{F}_n$-measurable. Consider the following limit order book $\tilde{L}$. The arrival process of $\tilde{L}$ is given by $(E_{t+n})_{t\ge 0}$; the starting state of $\tilde{L}$ is $(\tilde{\mathfrak{Q}}^b_0, \tilde{\mathfrak{Q}}^a_0) \equiv (\mathfrak{Q}^b_n, \mathfrak{Q}^a_n)$. Then at all times $t \ge 0$, $(\tilde{\mathfrak{Q}}^b_t, \tilde{\mathfrak{Q}}^a_t) = (\mathfrak{Q}^b_{t+n}, \mathfrak{Q}^a_{t+n})$. Consequently, $\tilde{\mathcal{A}}^b(x)$ holds for $\tilde{L}$ if and only $\mathcal{A}^b(x)$ holds for L.

Now consider a limit order book $\hat{L}$ with arrival process $(E_{t+n})_{t\ge 0}$ (same as for $\tilde{L}$) but starting state $(\mathfrak{Q}^b_0, \mathfrak{Q}^a_0)$ (same as for L). By construction, $\hat{\mathcal{A}}^b(x) \in \mathcal{F}_n$. On the other hand, since the starting states of $\tilde{L}$ and $\hat{L}$ differ by a finite number of orders, Corollary 3.15 implies that $\hat{\mathcal{A}}^b(x)$ holds for $\hat{L}$ if and only if $\tilde{\mathcal{A}}^b(x)$ holds for $\tilde{L}$, i.e. the events $\mathcal{A}^b(x)$ and $\hat{\mathcal{A}}^b(x)$ coincide.

We conclude that $\mathcal{A}^b(x)$ is $\mathcal{F}_n$-measurable for all n, and therefore it is $\mathcal{F}$-measurable. The argument above also demonstrates that whether $\mathcal{A}^b(x)$ holds for L is unaffected by finite changes in the starting state of L.

By Kolmogorov's 0-1 law [Williams, 1991, Theorem 4.11], for each x the event $\mathcal{A}^b(x)$ holds with probability 0 or 1. Let

$$\kappa_b \equiv \sup\{x : \mathbb{P}(\mathcal{A}^b(x)) = 1\}.$$

We claim that κ_b satisfies the conditions of the theorem.

First, for any $x < \kappa_b$ there exists $x \le y < \kappa_b$ such that $\mathbb{P}(\mathcal{A}^b(y)) = 1$. This implies $\mathbb{P}(\mathcal{A}^b(x)) = 1$ as well, since for $x \le y$ we have $\mathfrak{D}^b_\infty(-\infty, x] \le \mathfrak{D}^b_\infty(-\infty, y]$. Consequently, for any $\epsilon > 0$, only finitely many bids at prices below $\kappa_b - \epsilon$ ever depart the system, and similarly for asks.

It remains to show that, almost surely, $\mathfrak{Q}^b_t[\kappa_b + \epsilon, \infty) = 0$ infinitely often. This is by construction: for any $x > \kappa_b$ there will be infinitely many bid departures at prices $\le x$. At any time that a bid at price $\le x$ leaves the system, it must be the rightmost bid, so $\mathfrak{Q}^b_t(x, \infty) = 0$ happens infinitely often for any $x > \kappa_b$. □

We now consider the case of iid arrivals, and investigate the number of asks in the system at the times when there are no bids to the right of $\kappa_b + \epsilon$.

PROOF OF THEOREM 3.7. We know from Theorem 3.5 that κ_b and κ_a exist and are unique. The assertion $F^b(\kappa_b) = 1 - F^a(\kappa_a)$ is a consequence of the fact that the arrival distributions are absolutely continuous, and bids and asks always depart in pairs. Indeed, the long-run proportion of arriving bids that leave the system is $1 - F^b(\kappa_b)$, and this must equal the long-run proportion of arriving asks that leave the system, namely $F^a(\kappa_a)$.

Let $T_n \to \infty$ be the sequence of times along which $\mathfrak{Q}^b_{T_n}[\kappa_b + \epsilon, \infty) = 0$. We analyse the number of waiting asks $\mathfrak{Q}^a_{T_n}(-\infty, \kappa_a + \epsilon]$.

Using (83), we may write for any T,

$$\frac{1}{T}\mathfrak{Q}^a_T(-\infty, \kappa_a + \epsilon] = \frac{1}{T}\left(\mathfrak{A}^a_T(-\infty, \kappa_a + \epsilon] - \mathfrak{D}^a_T(-\infty, \kappa_a + \epsilon]\right).$$

Law of large numbers implies for the first term

$$\lim_{T\to\infty} \frac{1}{T}\mathfrak{A}^a_T(-\infty, \kappa_a + \epsilon] = \frac{1}{2}F^a(\kappa_a + \epsilon), \quad \text{w.p.1.}$$

For the second term,

$$\mathfrak{D}^a_T(-\infty, \kappa_a + \epsilon] = D^a_T(-\infty) - D^a_T(\kappa_a + \epsilon).$$

Theorem 3.5 implies that the second term is finite w.p.1. Since also $D^b_T(\kappa_b - \epsilon) < \infty$ w.p.1, we obtain

$$\lim_{T\to\infty} \frac{1}{T}\left(\mathfrak{D}^a_T(-\infty, \kappa_a + \epsilon] - \mathfrak{D}^b_T[\kappa_b - \epsilon, \infty)\right) = \lim_{T\to\infty} \frac{1}{T}\left(D^a_T(-\infty) - D^b_T(-\infty)\right) = 0, \quad \text{w.p.1}$$

since bids and asks always depart in pairs.

We would like to translate the statement about asks into a statement about bids. Observe

$$\mathfrak{D}^b_T[\kappa_b - \epsilon, \infty) \le \mathfrak{D}^b_T[\kappa_b + \epsilon, \infty) + \mathfrak{Q}^b_0[\kappa_b - \epsilon, \kappa_b + \epsilon) + \mathfrak{A}^b_T[\kappa_b - \epsilon, \kappa_b + \epsilon).$$

The second term is finite by assumption, while for the third term, w.p.1,

$$\lim_{T\to\infty} \frac{1}{T}\mathfrak{A}^b_T[\kappa_b - \epsilon, \kappa_b + \epsilon) = \frac{1}{2}\left(F^b(\kappa_b + \epsilon) - F^b(\kappa_b - \epsilon)\right) \le \epsilon \sup_{p\in[0,1]} f^b(p).$$

We conclude

$$\frac{1}{T}\mathfrak{Q}^a_T(-\infty, \kappa_a + \epsilon] \le \frac{1}{2}F^a(\kappa_a + \epsilon) - \frac{1}{T}\mathfrak{D}^b_T[\kappa_b + \epsilon, \infty) + \epsilon \sup_{p\in[0,1]} f^b(p) + o(1), \quad \text{w.p.1} \tag{86}$$

where the $o(1)$ term tends to 0 as $T \to \infty$.

We bound the first term as

$$F^a(\kappa_a + \epsilon) \le F^a(\kappa_a) + \epsilon \sup_{p\in[0,1]} f^a(p) \le F^b(\kappa_b + \epsilon) + \epsilon\left(\sup_{p\in[0,1]} f^a(p) + \sup_{p\in[0,1]} f^b(p)\right).$$

Recall $M = \max_{i=a,b} \sup_{p\in[0,1]} f^i(p)$. Putting the estimates above together, we conclude

$$\frac{1}{T}\mathfrak{Q}^a_T(-\infty, \kappa_a + \epsilon] \le \frac{1}{T}\left(\mathfrak{A}^b_T[\kappa_b + \epsilon, \infty) - \mathfrak{D}^b_T[\kappa_b + \epsilon, \infty)\right) + 2\epsilon M + o(1).$$

Finally, looking along the sequence T_n, the first term vanishes, so

$$\limsup_{T_n\to\infty} \frac{1}{T_n}\mathfrak{Q}^a_{T_n}(-\infty, \kappa_a + \epsilon] \le 2M\epsilon.$$

□

5. Strict limit order book

We would like to analyse the steady-state behaviour of the system when the pricing is continuous: that is, $\mathcal{P}(x) = x$. Lemma 3.17 allows us to bound the bid and ask counts $Q^b_t(x)$, $Q^a_t(x)$ for this case from below, but not from above. To be able to bound them from above, we introduce the model of a *strict limit order book*.

A strict limit order book differs from the ordinary one in that a bid at price p and an ask at price q are only allowed to depart the system when $q \prec p$ (i.e., the inequality must be strict). In particular, a given price level may contain both a waiting bid and a waiting ask. This does not affect the dynamics when $\mathcal{P}(x) = x$, since w.p.1 all orders in that system arrive at distinct prices. On the other hand, if for some price p the set $\{x : x \sim p\}$ has positive measure, then this modification will alter the paths of the limit order book, because fewer bid-ask pairs will be eligible to leave.

The main results in this section are Lemma 3.21 and Corollary 3.25. Lemma 3.21 shows that when the arrival process and the price level function are sufficiently "nice", strict and nonstrict limit order books are really quite similar. Corollary 3.25 concludes that we can bound the constant κ_b for a limit order book with continuous pricing from above *and* below using only ordinary limit order books with discrete pricing.

LEMMA 3.20. *Let L and $\tilde{L}$ be strict limit order books with the same starting state and external arrival process, but let $\tilde{\mathcal{P}}$ be coarser than $\mathcal{P}$ (Definition 3.16). Then*

$$\tilde{D}^b_t(p) \le D^b_t(p), \quad \tilde{D}^a_t(p) \le D^a_t(p), \quad \forall p, \forall t \ge 0.$$

The proof is similar to the proof of Lemma 3.17.

We now show that the strict and non-strict versions of the limit order book are not too different.

Let f^b_N, f^a_N be defined on counting measures of bids and asks as follows: for bids,

$$f^b_N(\mathfrak{Q}^b_t) = f^b_N(\sum b_i\delta_{p_i}) \equiv \sum \mathbf{1}_{p_i \le 1-\frac{1}{N}} b_i\delta_{p_i};$$

$$f^a_N(\mathfrak{Q}^a_t) = f^a_N(\sum a_i\delta_{p_i}) \equiv \sum \mathbf{1}_{p_i \ge \frac{1}{N}} a_i\delta_{p_i-\frac{1}{N}}.$$

LEMMA 3.21. *Let L be a limit order book with empty starting state, arrival process $(A_t)_{t\ge0}$ with arrivals iid uniform on $[0,1] \times \{bid, ask\}$ and price level function $\mathcal{P}(x) = \frac{1}{N}\lfloor Nx \rfloor$. Let $\check{L}$ be a limit order book with empty starting state, the same price level function $\mathcal{P}$, and arrival events $\check{A}_t$ given as follows. If $A_t = (p, ask)$ for $p > \frac{1}{N}$, then $\check{A}_t = (p - \frac{1}{N}, ask)$. If $A_t = (p, bid)$ for $p < 1 - \frac{1}{N}$, then $\check{A}_t = (p, bid)$. (The remaining events are ignored.)*

Then the state of $\check{L}$ at time t is given by

$$\check{\mathfrak{Q}}^b_t = f^b_N(\mathfrak{Q}^b_t), \quad \check{\mathfrak{Q}}^a_t = f^a_N(\mathfrak{Q}^a_t).$$

PROOF. An easy induction on t. □

We derive the following two corollaries for Poisson arrival processes.

COROLLARY 3.22. *Let L be a limit order book with empty starting state, price level function $\mathcal{P}(x) = \frac{1}{N}\lfloor Nx \rfloor$, and arrival process which is Poisson on the set $[0,1]\times\{bid, ask\}\times[0,\infty)$ (the last coordinate being time), where arrival events have rate 1 in time and are uniform on $[0,1]\times\{bid, ask\}$. We work on the probability-one event that the set of times at which arrivals occur is discrete, and times of arrival events do not coincide. Let $\check{L}$ be a limit order book with empty starting state, price level function $\check{\mathcal{P}}(x) = \frac{1}{N-1}\lfloor (N-1)x \rfloor$, and arrival process which is Poisson on the set $[0,1]\times\{bid, ask\}\times[0,\infty)$ (the last coordinate being time), where arrival events have rate $\frac{N-1}{N}$ in time and are uniform on $[0,1]\times\{bid, ask\}$.*

For a counting measure $\sum c_i\delta_{p_i}$ with support on $[0, \frac{N-1}{N}]$ define

$$\tilde{f}(\sum c_i\delta_{p_i}) = \sum c_i\delta_{\frac{N}{N-1}p_i}.$$

Then $\check{\mathfrak{Q}}^i_t$ and $\tilde{f}\big(f^i_N(\mathfrak{Q}^i_t)\big)$, $i = a, b$, are equal in distribution.

PROOF. Lemma 3.21 gives a coupling. Observe that the construction given in the lemma applied to a Poisson process of rate 1 for the ordinary limit order book produces an arrival process that is Poisson of rate $\frac{N-1}{N}$ in time for the strict limit order book, and which is uniform on $[0, \frac{N-1}{N}]$. If we want the arrival events to be uniform on $[0,1]$ instead, we simply need to rescale space using $\tilde{f}$. □

COROLLARY 3.23. *Let L be a limit order book as in Corollary 3.22. Let $\check{L}$ be a limit order book with empty starting state, price level function $\check{\mathcal{P}}(x) = \frac{1}{N-1}\lfloor (N-1)x \rfloor$, and arrival process which is Poisson on the set $[0,1]\times\{bid, ask\}\times[0,\infty)$ (the last coordinate being time), where arrival events have rate 1 in time and are uniform on $[0,1]\times\{bid, ask\}$. We work on the probability-one event that the set of times at which arrivals occur is discrete, and times of arrival events do not coincide.*

Then $\check{\mathfrak{Q}}^i_{\frac{N-1}{N}t}$ and $\tilde{f}\big(f^i_N(\mathfrak{Q}^i_t)\big)$, $i = a, b$, are equal in distribution.

PROOF. For a Poisson process, rescaling time is equivalent to rescaling the rate, so this follows from Corollary 3.22. □

COROLLARY 3.24. *Let L be an ordinary limit order book with finite starting state, price level function $\mathcal{P}(x) = \frac{1}{N}\lfloor Nx \rfloor$, and let the arrival events $(A_t)_{t\geq 0}$ be iid uniform. Let $\check{L}$ be a strict limit order book with finite starting state, price level function $\check{\mathcal{P}}(x) = \frac{1}{N-1}\lfloor (N-1)x \rfloor$, and the same arrival process. Then*

$$\frac{N}{N-1}\kappa_b = \check{\kappa}_b, \quad \frac{N}{N-1}(1-\kappa_a) = 1-\check{\kappa}_a.$$

where κ_i (with or without the hat) are defined as in Theorem 3.5.

PROOF. Since the statements about κ_i do not depend on the time scaling of the arrival process so long as the arrival events are independent, we may take the arrival process to be Poisson in time. Further, as we saw in the proof of Theorem 3.5, the statements about κ do not depend on the starting state provided it's finite. Applying Corollary 3.23 we conclude that the defining properties of Theorem 3.5 are simultaneously true or not true of κ_b (respectively $1-\kappa_a$) in the ordinary system, and of $\check{\kappa}_b \equiv \frac{N}{N-1}\kappa_b$ (respectively $1-\check{\kappa}_a \equiv \frac{N}{N-1}(1-\kappa_a)$) in the strict system. □

Combining Corollary 3.24 with Lemmas 3.17 and 3.20 gives the following

COROLLARY 3.25. *Consider the limit order book with finite starting state, $\mathcal{P}(x) = x$, and arrival events $(A_t)_{t\geq 0}$ iid uniform on $[0,1]\times\{bid, ask\}$. Let κ_i, $i = a, b$ be the associated threshold values (Definition 3.6). Consider also the ordinary limit order books*

L^N and strict limit order books $\check{L}^N$, with the same arrival process and $\mathcal{P}^N(x) = \frac{1}{N}\lfloor Nx \rfloor$, $\check{\mathcal{P}}^N(x) = \frac{1}{N-1}\lfloor (N-1)x \rfloor$. Let κ_i^N, $\check{\kappa}_i^N$, $i = a, b$ be the corresponding threshold values. Then

$$\kappa_b^N \leq \kappa_b \leq \check{\kappa}_b^N = \frac{N}{N-1}\kappa_b^N, \tag{87}$$

$$1 - \kappa_a^N \leq 1 - \kappa_a \leq 1 - \check{\kappa}_a^N = \frac{N}{N-1}(1 - \kappa_a^N). \tag{88}$$

6. Exact values of the thresholds κ_b and κ_a

In this section we will derive the value of κ_b for the case of iid uniform arrivals and price level function $\mathcal{P}(x) = x$. In the process, we will conjecture steady-state distributions for the rightmost bid and the leftmost ask. We will find certain sequences of times with the property that empirical distributions converge to the conjectured distributions.

For the duration of this section, we will assume that the arrival events are iid uniform on $[0, 1] \times \{\text{bid}, \text{ask}\}$. We will consider price level functions with a finite number of distinct price levels. Typically, but not necessarily, we will be interested in $\mathcal{P}(x) = \frac{1}{N}\lfloor Nx \rfloor$ for some N. In all the models we investigate, the price level functions will satisfy $x \prec y$ if and only if $1 - y \prec 1 - x$; together with the symmetry of the arrival process about the point $\frac{1}{2}$, this implies that the threshold values satisfy $\kappa_b = 1 - \kappa_a$.

Definition 3.26. We will refer to the set of prices in the k^{th} price level as *bin* k. Let l_k be the width of the k^{th} bin; that is, the k^{th} bin is

$$[l_1 + \ldots + l_{k-1}, l_1 + \ldots + l_k) = \{x : x \sim l_1 + \ldots + l_{k-1}\}$$

(Recall that $\mathcal{P}$ was taken to be right-continuous in Definition 3.3.) For $x \in [0, 1]$, we let $[x]$ denote the index of the bin to which x belongs; that is, $[x] = k$ if x belongs to bin k.

Write $\kappa_b(N)$ and $\kappa_a(N)$ for the constants identified in Theorem 3.5. Let $k_b(N) = [\kappa_b(N)]$, unless $\kappa_b = l_1 + \ldots + l_{k-1}$ happens to coincide with the leftmost boundary of bin k; in that case, define $k_b(N) = k - 1$. Similarly, let $k_a(N) = [\kappa_a(N)]$ (since we are taking bins to be closed on the left and open on the right, the corresponding caveat is unnecessary).

For a measure π on $[0, 1]$ we abuse notation slightly, and write write $\pi(k)$ to mean the measure of bin k.

Theorem 3.5 implies that the number of bid departures from bins $< k_b(N)$, and ask departures from bins $> k_a(N)$, will be finite.

For each $\delta > 0$ let $T_n = T_n(\delta) \to \infty$ be the sequence of (random) times along which

$$\mathfrak{D}^b_{T_1}(-\infty, \kappa_b(N) - \delta] = \mathfrak{D}^b_{\infty}(-\infty, \kappa_b(N) - \delta], \text{ and} \tag{89a}$$

$$\mathfrak{D}^a_{T_1}[\kappa_a(N) + \delta, \infty) = \mathfrak{D}^a_{\infty}[\kappa_a(N) + \delta, \infty); \tag{89b}$$

$$\mathfrak{Q}^b_{T_n}[\kappa_b(N) + \delta, \infty) = 0, \text{ and} \tag{89c}$$

$$\frac{1}{T_n}\mathfrak{Q}^a_{T_n}(-\infty, \kappa_a(N) + \delta] \leq \delta. \tag{89d}$$

The (almost-sure) existence of such a sequence has been established in Theorems 3.5 and 3.7.

For any time T and $1 \leq k \leq N$, we may define the empirical distributions of the rightmost bid β_t and the leftmost ask α_t up to time T as follows:

$$\pi_T^b(k) = \frac{1}{T}\#\left\{0 \leq t \leq T : [\beta(t)] = k\right\},$$
$$\pi_T^a(k) = \frac{1}{T}\#\left\{0 \leq t \leq T : [\alpha(t)] = k\right\}$$

These are discrete probability distributions on the compact set $[0, 1]$; consequently, there is some subsequence of our chosen sequence $T_n(\delta)$ (which, by a slight abuse of notation, we also call T_n) along which we have convergence

$$\pi_{T_n}^b \xrightarrow{w} \pi^b, \quad \pi_{T_n}^a \xrightarrow{w} \pi^a$$

for some discrete probability measures π^b and π^a. We will now use the relations in (89) to derive some properties of these limiting distributions π^a and π^b.

Remark 3.27. Slightly more is true than stated above. Namely, *any* subsequence of $T_n(\delta)$ has a further subsequence, say T_{n_r}, along which the $\pi_{T_{n_r}}^b$ converge to some limit, and this limit will have all the properties given below.

Lemma 3.28. *With the above definitions,*

$$\left(\sum_{i \leq k} l_i\right) \frac{\pi^b(k)}{l_k} + \sum_{i \leq k} \pi^a(i) = 1$$

and

$$1 - \frac{\delta}{\min l_i} \leq \left(\sum_{i=k}^{N} l_i\right) \frac{\pi^a(k)}{l_k} + \sum_{i \leq k} \pi^b(i) \leq 1$$

Proof. Note that, for any set S, we have

$$\mathfrak{Q}_T^b(S) - \mathfrak{Q}_0^b(S) = \sum_{t \leq T} \mathbf{1}_{\{A_t = p_t \times \text{bid},\, p_t \in S,\, \alpha_t \succ p_t\}} - \sum_{t \leq T} \mathbf{1}_{\{A_t = p_t \times \text{ask},\, \beta_t \in S,\, p_t \not\succ \beta_t\}}$$

We now estimate these two terms when S is bin k.

The first term is the number of exogenous bid arrivals, up to time T, during the times when $[\alpha_t] > k$. Recalling that exogenous arrivals are independent of the state of the limit order book,

$$\sum_{t \leq T} \mathbf{1}_{\{A_t = p_t \times \text{bid},\, p_t \in S,\, \alpha_t \succ p_t\}} = \text{Bin}\left(\frac{1}{2} l_k \Big(\sum_{i>k} \pi_T^a(i)\Big) T\right)$$

is a binomial random variable with parameter $\frac{1}{2} l_k (\sum_{i>k} \pi_T^a(i)) T$. Indeed, during each of the $\sum_{i>k} \pi_T^a(i) T$ times when $[\alpha_t] > k$ there is a probability $\frac{1}{2} l_k$ of the arrival event being a bid at a price $p_t \in S$.

The second term is the number of external ask arrivals, up to time T, into bins $\leq k$, during the times when $[\beta_t] = k$. We obtain

$$\sum_{t \leq T} \mathbf{1}_{\{A_t = p_t \times \text{ask},\, \beta_t \in S,\, p_t \not\succ \beta_t\}} = \text{Bin}\left(\frac{1}{2} \Big(\sum_{i \leq k} l_i\Big) \pi_T^b(k) T\right)$$

is a binomial random variable with parameter $\frac{1}{2} (\sum_{i \leq k} l_i) \pi_T^b(k) T$, since during each of the $\pi_T^b(k) T$ times when $[\beta_t] = k$, the probability of an ask arriving at price $p_t \not\succ l_1 + \ldots + l_k$ is $(\sum_{i \leq k} l_i)$.

Dividing by T, looking along T_n, and using (89c) to evaluate the left-hand side, we obtain

$$0 = \frac{1}{T_n}\mathrm{Bin}\left(\frac{1}{2}l_k\Big(\sum_{i>k}\pi^a_{T_n}(i)\Big)T_n\right) - \frac{1}{T_n}\mathrm{Bin}\left(\frac{1}{2}\Big(\sum_{i\le k}l_i\Big)\pi^b_{T_n}(k)T_n\right). \tag{90}$$

By the law of large numbers, this implies

$$0 = \lim_{n\to\infty}\left(\frac{1}{2}l_k\big(\sum_{i>k}\pi^a_{T_n}(i)\big) - \frac{1}{2}\big(\sum_{i\le k}l_i\big)\pi^b_{T_n}(k)\right).$$

Using the fact that $\sum_i \pi^a_{T_n}(i) = 1$ for all n, and $\pi^i_{T_n} \xrightarrow{w} \pi^i$ for $i = a, b$, we conclude that

$$\big(\sum_{i\le k}l_i\big)\frac{\pi^b(k)}{l_k} + \sum_{i\le k}\pi^a(i) = 1.$$

The equation for asks is obtained similarly, using (89d) instead of (89c) in (90). □

Observe that if $\lim_{n\to\infty}\frac{1}{T_n}\mathfrak{Q}^b_{T_n}(S) = Q$ exists and is well-defined, (90) becomes

$$Q = \frac{1}{T_n}\mathrm{Bin}\left(\frac{1}{2}l_k\big(\sum_{i>k}\pi^a_{T_n}(i)\big)T_n\right) - \frac{1}{T_n}\mathrm{Bin}\left(\frac{1}{2}\big(\sum_{i\le k}l_i\big)\pi^b_{T_n}(k)T_n\right),$$

and in the limit we obtain

$$\big(\sum_{i\le k}l_i\big)\frac{\pi^b(k)}{l_k} + \sum_{i\le k}\pi^a(i) = 1 - \frac{2Q}{l_k}. \tag{91}$$

We can now prove Theorem 3.9.

Proof of Theorem 3.9. We begin by analysing a model with three bins of width l, $1-2l$, and l again; that is, $N = 3$ in the above machinery. We will see how this relates to the original limit order book later.

By symmetry, $\kappa_a = 1 - \kappa_b$; the two clearly cannot be in the middle bin, so $k_b = 1$ and $k_a = 3$. Let $\delta > 0$ be small. We will consider the sequence of times T_n along which (89) holds. We deduce the following variants of (90):

$$\pi^b(1) + \pi^a(1) = 1 - \frac{\kappa_b}{l} \tag{92a}$$

$$\pi^a(3) + \pi^b(3) \in \left[1 - \frac{\kappa_b + \delta}{l}, 1 - \frac{\kappa_b}{l}\right] \tag{92b}$$

$$\frac{1-l}{1-2l}\pi^b(2) + \pi^a(2) + \pi^a(1) = 1 \tag{92c}$$

$$\frac{1-l}{1-2l}\pi^a(2) + \pi^b(2) + \pi^b(3) \in \left[1 - \frac{\delta}{l}, 1\right] \tag{92d}$$

$$\frac{1}{l}\pi^b(3) + \pi^a(1) + \pi^a(2) + \pi^a(3) = 1 \tag{92e}$$

$$\frac{1}{l}\pi^a(1) + \pi^b(1) + \pi^b(2) + \pi^b(3) \in \left[1 - \frac{\delta}{l}, 1\right] \tag{92f}$$

(Note that $\lim_{t\to\infty}\frac{1}{t}\mathfrak{Q}^b_t[0,l) = \frac{1}{2}\kappa_b$, and similarly for asks in $[1-l, 1]$.) Taking

$$(2-3l)((92a) + (92b)) + (1-2l)((92c) + (92d)) - l(1-2l)((92e) + (92f))$$

yields

$$(2-4l+2l^2)(\pi^b(1)+\pi^a(1)+\pi^b(2)+\pi^a(2)+\pi^b(3)+\pi^a(3)) \in \\ 6-12l+4l^2-(\frac{4}{l}-6)\kappa_b+\left[-\frac{3-5l}{l}\delta,(1-2l)\delta\right].$$

The left-hand side is clearly $\leq 2(2-4l+2l^2)$. Therefore,

$$\kappa_b \geq \frac{l-2l^2}{2-3l}+C\delta$$

for some constant C. Since δ can be chosen arbitrarily small, we conclude that $\kappa_b \geq \frac{l-2l^2}{2-3l}$ for the three-bin model.

The result of the theorem now follows by applying Corollary 3.19. Indeed, for any limit order book L, let p be such that $0 \prec p \prec 1-p \prec 1$, and consider the limit order book $\tilde{L}$ with the coarser pricing scheme

$$\tilde{\mathcal{P}}(x)=\begin{cases}0, & x \leq p \\ \frac{1}{2}, & p<x\leq 1-p \\ 1, & 1-p<x\end{cases}$$

Corollary 3.19 implies that $\frac{p-2p^2}{2-3p} \leq \tilde{\kappa}_b \leq \kappa_b$. In particular, if $\mathcal{P}(x)=x$ (so p can be chosen arbitrarily), then choosing $p=1/3$ maximises this bound at $\frac{1}{9}$. □

We now establish some further properties of the limiting distributions π^a and π^b, for arbitrary $N \geq 3$.

LEMMA 3.29. (1) *$\pi^b(k)=0$ for $k \leq k_b(N)-1$ or $k \geq k_a(N)$. Similarly, $\pi^a(k)=0$ for $k \geq k_a(N)+1$ or $k \leq k_b(N)$.*

(2) *$\frac{\pi^b(k)}{l_k} \leq \left(\sum_{i\leq k} l_i\right)^{-1}$, and $\frac{\pi^a(k)}{l_k} \leq \left(\sum_{i\geq k} l_i\right)^{-1}$, for all $k=1,\ldots,N$.*

(3)

$$\frac{\pi^b(k_a(N)-1)}{l_{k_a(N)-1}} \leq l_{k_a(N)-1}\left(\sum_{i\geq k_a(N)-1} l_i\right)^{-1}\left(\sum_{i\leq k_a(N)-1} l_i\right)^{-1}$$

and

$$\frac{\pi^a(k_b(N)+1)}{l_{k_b(N)+1}} \leq l_{k_b(N)+1}\left(\sum_{i\leq k_b(N)+1} l_i\right)^{-1}\left(\sum_{i\geq k_b(N)+1} l_i\right)^{-1}.$$

PROOF. For (1), note that if for some k we have $\pi^b(k)>0$, then for some $\epsilon>0$ and all sufficiently large n we must have $\pi^b_n(k)>\epsilon$, i.e. $\#\{t \leq T_n : [\beta(t)]=k\} > \epsilon T_n$ for all large n. In that case law of large numbers implies

$$\liminf_{n\to\infty} \frac{1}{T_n}\mathfrak{D}^b_{T_n}(\text{bin } k) > \epsilon.$$

Since for $k \leq k_b(N)-1$ this limit is equal to 0, we conclude $\pi^b(k)=0$ for $k \leq k_b(N)-1$. By an identical argument, $\pi^a(k)=0$ for $k \geq k_a(N)+1$. Further, after a finite time there are always asks in bin $k_a(N)$. Consequently, after a finite time the rightmost bid cannot be in any bin $k \geq k_a(N)$, so $\pi^b(k)=0$ for $k \geq k_a(N)$ By an identical argument, $\pi^a(k)=0$ for $k \leq k_b(N)$.

For (2), define $A_T^{b,Q}(k)$ (respectively $D_T^{b,Q}(k)$) to be the number of bids arriving into (respectively departing from) queue in bin k up to time T (as opposed to departing without ever visiting the queue). Observe

$$D_T^{b,Q}(k) = \text{Bin}\left(\frac{1}{2}\pi_T^b(k)T\big(\sum_{i\le k} l_i\big)\right).$$

The number of arrivals into queue in bin k is certainly bounded above by the total number of arrivals into bin k, namely

$$A_T^{b,Q}(k) \le \mathfrak{A}_T^b(k) = \text{Bin}\big(\frac{1}{2}l_k T\big).$$

Dividing by T and looking along $T_n \to \infty$,

$$\lim_{n\to\infty}\frac{1}{T_n}D_T^{b,Q}(k) = \frac{1}{2}\pi^b(k)\big(\sum_{i\le k} l_i\big)$$

and

$$\lim_{n\to\infty}\frac{1}{T_n}\mathfrak{A}_{T_n}^b(k) = \frac{1}{2}l_k.$$

The inequality "number of departures $\le$ number of arrivals" thus gives $\pi^b(k)(\sum_{i\le k} l_i) \le l_k$. By an identical argument for asks, we conclude

$$\frac{\pi^b(k)}{l_k} \le \big(\sum_{i\le k} l_i\big)^{-1}, \quad \frac{\pi^a(k)}{l_k} \le \big(\sum_{i\ge k} l_i\big)^{-1}, \quad \forall k.$$

Finally, for (3), we use the definition of the time sequence T_n to refine the computation on the number of arrivals into queue. Note that external arrivals into bin k go into queue if only if $[\alpha_t] > k$: that is, we have

$$E_T^{b,Q}(k) = \text{Bin}\left(\frac{1}{2}\big(\sum_{i>k}\pi_T^a(i)\big)Tl_k\right).$$

Let $T_n \to \infty$, and recall by (1)

$$\pi^a(k) = 0,\ k \ge k_a(N)+1, \quad \pi^a(k_a(N)) \le l_k\left(\sum_{i\ge k} l_i\right)^{-1}.$$

This gives

$$\limsup_{n\to\infty}\frac{1}{T_n}E_{T_n}^{b,Q}(k_a(N)-1) \le \frac{1}{2}l_{k_a(N)-1}^2\left(\sum_{i\ge k_a(N)-1} l_i\right)^{-1};$$

while for the departures we still have

$$\lim_{n\to\infty}\frac{1}{T_n}D_{T_n}^{b,Q}(k_a(N)-1) = \frac{1}{2}\pi^b(k_a(N)-1)\left(\sum_{i\le k_a(N)-1} l_i\right).$$

Combining the two yields the inequality

$$\frac{\pi^b(k_a(N)-1)}{l_{k_a(N)-1}} \le l_{k_a(N)-1}\left(\sum_{i\ge k_a(N)-1} l_i\right)^{-1}\left(\sum_{i\le k_a(N)-1} l_i\right)^{-1},$$

and similarly for asks. □

We now let $N \to \infty$:

PROPOSITION 3.30. (1) *Let $T_n = T_n(\delta)$ be a sequence along which (89) is satisfied, and also $\pi^b_{T_n} \xrightarrow{w} \pi^b$, $\pi^a_{T_n} \xrightarrow{w} \pi^a$ for some distributions π^b, π^a. Then the limiting distributions satisfy*

$$\text{(93a)} \qquad \big(\sum_{i\le k} l_i\big)\frac{\pi^b(k)}{l_k} + \sum_{i\le k}\pi^a(i) = 1, \quad \sum_{k=k_b(N)+1}^{k_a(N)} \pi^b(k) \ge 1 - l_{k_b(N)}\left(\sum_{i\le k_b(N)} l_i\right)^{-1}$$

and

$$\text{(93b)} \quad 1 - \frac{\delta}{\min l_i} \le \big(\sum_{i=k}^{N} l_i\big)\frac{\pi^a(k)}{l_k} + \sum_{i\le k}\pi^b(i) \le 1,$$

$$\sum_{k=k_b(N)}^{k_a(N)-1} \pi^a(k) \ge 1 - l_{k_a(N)}\left(\sum_{i\ge k_a(N)} l_i\right)^{-1}.$$

(2) *Consider a sequence of limit order books indexed by $N = 1, 2, \ldots$, such that the partial orderings $\mathcal{P}^N$ are symmetric and are successive refinements. That is, $x \prec^N y$ if and only if $1 - y \prec^N 1 - x$ for all N, and if $x \prec^N y$ then $x \prec^n y$ for all $n \ge N$. We further require*

$$\max_{1\le i\le N} l_i \to 0, \quad \frac{\delta_N}{\min l_i} \to 0, \qquad \text{as } N \to \infty.$$

Let $T^N_n(\delta_N)$ be the corresponding family of sequences along which (89) is satisfied. Each sequence of distributions $\pi^{i,N}_{T^N_n}$, $i = a, b$, has a convergent subsequence; call its limit $\pi^{i,N}$. (There may be multiple possible values for $\pi^{i,N}$, depending on the convergent subsequence we choose; pick any.) As $N \to \infty$, for some functions ϖ^i, the following pointwise convergence holds:

$$\frac{1}{l_{[x]}}\pi^{b,N}([x]) \to \varpi^b(x), \quad \frac{1}{l_{[x]}}\pi^{a,N}([x]) \to \varpi^a(x).$$

The limits ϖ^b, ϖ^a satisfy

$$\text{(94)} \qquad \varpi^b(x) = \mathbf{1}_{(\kappa,1-\kappa)}(1-\kappa)\left(\frac{1}{x} + \log(\frac{1-x}{x})\right), \quad \varpi^a(x) = \varpi^b(1-x),$$

where κ is the unique real number satisfying

$$\log\frac{1-\kappa}{\kappa} = \frac{1}{1-\kappa}, \quad \kappa \approx 0.2178.$$

Before proving this result, we derive Theorem 3.11 from it:

COROLLARY (Theorem 3.11). *Let the partial ordering $\mathcal{P}$ be $\mathcal{P}(x) = x$, i.e. $x \prec y$ if and only if $x < y$. Let the arrival events be iid uniform on $[0,1] \times \{bid, ask\}$. Then the value of κ_b is given as the unique solution to*

$$\log\left(\frac{1-\kappa_b}{\kappa_b}\right) = \frac{\kappa_b}{1-\kappa_b} + 1, \quad \kappa_b \approx 0.2178$$

and $\kappa_a = 1 - \kappa_b$.

PROOF. We combine Proposition 3.30 with Corollary 3.25, noting that

$$\mathcal{P}^N(x) = \frac{1}{N!}\lfloor N!x\rfloor, \quad \delta_N = (N!)^{-2}$$

clearly satisfies the conditions of Proposition 3.30. □

Proof of Proposition 3.30. (1) simply combines Lemma 3.28 with Lemma 3.29.

For (2), since the space of probability distributions on the compact set $[0,1]$ is compact, any sequence of distributions has a convergent subsequence, so there is no problem with defining some set of limits $\pi^{i,N}$ $(i=a,b)$, where $\pi^{i,N} = \lim_{n_k\to\infty} \pi^{i,N}_{T^N_{n_k}}$. Since these are again probability distributions on a compact set, any subsequence (along N_k, say) will have a further, convergent, subsequence. Below, we will show that the only possible limit of a convergent subsequence is the measure with density p^i given by (94); this proves the result.

We now work along a subsequence of N along which weak limits of the two sequences $\pi^{i,N}$ $(i=a,b)$ exist; by a slight abuse of notation, we still denote its index N. Rewrite the inequalities (93) as follows:

$$\text{(95a)}\qquad \Big(\sum_{i\le k} l_i\Big)\frac{\pi^{b,N}(k)}{l_k} + \sum_{i\le k} l_i\left(\frac{\pi^{a,N}(i)}{l_i}\right) = 1,\ \sum_{k=k_b(N)+1}^{k_a(N)} l_k\left(\frac{\pi^{b,N}(k)}{l_k}\right) = 1-\epsilon^{b,N}$$

and

$$\text{(95b)}\quad \Big(\sum_{i\ge k} l_i\Big)\frac{\pi^{a,N}(k)}{l_k} + \sum_{i\le k} l_i\left(\frac{\pi^{b,N}(i)}{l_i}\right) = 1-\epsilon^{a,N}(k),$$

$$\sum_{k=k_b(N)}^{k_a(N)-1} l_k\left(\frac{\pi^{a,N}(k)}{l_k}\right) = 1-\epsilon^{a,N}$$

for some constants and functions $\epsilon^{b,N}$, $\epsilon^{a,N}$, and $\epsilon^{a,N}(\cdot)$ which converge to 0 as $\max(\frac{\delta_N}{\min l_i}, l_i)\to 0$.

Lemma 3.29 (1) implies $\pi^{a,N}(k)=0$ for $k\le k_b(N)$. Inserting this into (95a) gives

$$\Big(\sum_{i\le k_b(N)+1} l_i\Big)\frac{\pi^{b,N}(k_b(N)+1)}{l_{k_b(N)+1}} = 1-\pi^{a,N}(k_b(N)+1) \equiv 1-\varepsilon_1^{b,N}$$

and

$$\Big(\sum_{i\le k_b(N)+2} l_i\Big)\frac{\pi^{b,N}(k_b(N)+2)}{l_{k_b(N)+2}} = 1-\pi^{a,N}(k_b(N)+1)-\pi^{a,N}(k_b(N)+2) \equiv 1-\varepsilon_2^{b,N}.$$

By Lemma 3.29 (3),

$$\pi^{a,N}(k_b(N)+1) \le l^2_{k_b(N)+1}\left(\sum_{i\le k_b(N)+1} l_i\right)^{-1}\left(\sum_{i\ge k_b(N)+1} l_i\right)^{-1}.$$

A similar inequality may be derived for $\pi^{a,N}_{k_b(N)+2}$. These inequalities imply that $\varepsilon^{b,N}_{1,2}\to 0$ as $\max l_i\to 0$ (Theorem 3.9 applies, so κ_b is bounded away from 0). Rearranging, we conclude

$$\text{(96a)}\qquad \left|\frac{\pi^{b,N}(k_b(N)+1)}{l_{k_b(N)+1}} - \left(\frac{\sum_{i\le k_b(N)+1} l_i}{N}\right)^{-1}\right| \to 0$$

and

$$\text{(96b)}\qquad \left|\left(\frac{\pi^{b,N}(k_b(N)+1)}{l_{k_b(N)+1}} - \frac{\pi^{b,N}_{k_b(N)+2}}{l_{k_b(N)+2}}\right) - \left(\frac{\sum_{i\le k_b(N)+1} l_i}{N}\right)^{-2} l_{k_b(N)+2}\right| \to 0.$$

Consider now the following system of integral equations:

$$x\varpi^b(x) + \int_0^x \varpi^a(y)dy = 1, \quad \int_\kappa^{1-\kappa} \varpi^b(x)dx = 1,$$
$$(1-x)\varpi^a(x) + \int_x^1 \varpi^b(y)dy = 1, \quad \int_\kappa^{1-\kappa} \varpi^a(x)dx = 1.$$

Here,

$$\kappa \equiv \lim \frac{k_b(N)}{N};$$

the limit exists because $\mathcal{P}^N$ are refinements, and hence Corollary 3.19 implies that $\kappa_b(N)$ is an increasing sequence which must have a limit. (The rounding error in going from $\kappa_b(N)$ to $\frac{k_b(N)}{N}$ will not cause a problem.)

We can rewrite these equations as follows: on $(\kappa, 1-\kappa)$ we have

$$\frac{d}{dx}(x\varpi^b(x)) = -\varpi^a(x) = -\frac{1}{1-x}\left(1 - \int_x^1 \varpi^b(y)dy\right) = -\frac{1}{1-x}\int_0^x \varpi^b(y)dy$$

and hence

$$\frac{d}{dx}\left(-(1-x)\frac{d}{dx}(x\varpi^b(x))\right) = \varpi^b(x). \tag{98a}$$

This is a second-order ordinary differential equation, which needs two initial conditions to have a unique solution. We take

$$\varpi^b(\kappa) = \frac{1}{\kappa}, \quad \frac{d}{dx}\varpi^b(x)|_{x=\kappa} = -\frac{1}{\kappa^2}. \tag{98b}$$

Now, the set of coupled first-order equations (95) can be rearranged to form a second-order difference equation for $\frac{\pi^{b,N}([x])}{l_{[x]}}$, with initial conditions given by (96). In this setting, results on convergence of Euler's method for numerically approximating ordinary differential equations [Bradie, 2006, Chapter 7] guarantee that the functions $\frac{1}{l_{[x]}}\pi^{i,N}([x])$ converge to the unique solution of the ODE (98a) with initial conditions (98b).

The ordinary differential equation (98) can be solved explicitly to give

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

Moreover, Lemma 3.29 implies that $\varpi^b(x) \to 0$ as $x \to 1-\kappa$. Putting this into the above equation gives

$$\log\frac{1-\kappa}{\kappa} = \frac{1}{1-\kappa}, \quad \kappa \approx 0.2178.$$

The claim $\varpi^a(x) = \varpi^b(1-x)$ follows easily from the symmetry of the system. □

The distribution $\varpi^b(x)$, together with simulated data for it, is plotted in Figure 3.2 (§11).

7. Restricted limit order book, and conjecture on steady-state behaviour

In §6 we have constructed distributions $\varpi^b(x)$, $\varpi^a(x)$ as limits, along a certain carefully chosen sequence, of the empirical distributions of the location of the rightmost bid (respectively leftmost ask) for a limit order book with a finite number of price bins. While this is enough to prove that the value of κ_b in a system with $\mathcal{P}(x) = x$ is ≈ 0.2178, it by no means proves anything about the steady-state distribution of the rightmost bid in that limit order book. In fact, it is not clear that it makes sense to speak of the steady-state distribution of the rightmost bid in an ordinary limit order book, because

whenever $\kappa_b > 0$, the limit order book is an obviously-transient Markov process. (While this does not prevent some marginal of the state, such as the rightmost bid, from having a steady-state distribution, it makes its existence nonobvious.)

In this section we construct a modified object, the *restricted limit order book*, which we conjecture to be a positive recurrent Markov process. We also prove that, in a certain sense, the ordinary limit order book and the restricted one behave in the same way.

DEFINITION 3.31. Let L be a limit order book with initial state L_0. For an interval $I \equiv [x, y]$ define the restriction of L to I, denoted L^I, as follows. The initial state is given by

$$(\tilde{\mathfrak{Q}}_0^{b,I}, \tilde{\mathfrak{Q}}_0^{a,I}) \equiv (\infty\delta_x + \mathbf{1}_I\mathfrak{Q}_0^b, \infty\delta_y + \mathbf{1}_I\mathfrak{Q}_0^a). \tag{99}$$

That is, we restrict the measures to I by multiplying them by the indicator function $\mathbf{1}_I$, and add infinitely many bids at the point x, and infinitely many asks at the point y. The price level function and arrival process in L^I are the same as in L. The *state* of the restricted limit order book is

$$(\mathfrak{Q}_t^{b,I}, \mathfrak{Q}_t^{a,I}) \equiv \mathbf{1}_{(x,y)}(\tilde{\mathfrak{Q}}_t^{b,I}, \tilde{\mathfrak{Q}}_t^{a,I});$$

this is a Markovian descriptor, because $\mathfrak{Q}_t^b\{x\} = \mathfrak{Q}_t^a\{y\} = \infty$ at all times t. Arrival events A_t of the form $p \times$ bid for $p \le x$, and of the form $p \times$ ask for $p \ge y$, do not change the state of the restricted limit order book.

Restrictions can be combined: $(L^I)^J = L^{I\cap J}$. For all of the examples we have considered so far, L coincides with $L^{[0,1]}$.

The restricted order book can be thought of as a modification of the price level function:

LEMMA 3.32. *Let L be an ordinary limit order book with price level function $\mathcal{P}$. Let L^I be the restriction of L to an interval $I \equiv [x, y]$. Let $\tilde{\mathcal{P}}$ be the price level function given by*

$$\tilde{\mathcal{P}}(p) = \begin{cases} 0, & p \le x \\ \mathcal{P}(p), & x < p < y \\ 1, & p \ge y \end{cases}$$

and let $\tilde{L}$ be an ordinary limit order book with price level function $\tilde{\mathcal{P}}$ and the same initial state and arrival process as L. Then at all times

$$\mathbf{1}_I\tilde{L}_t = L_t^I.$$

PROOF. Easy induction in t. □

In particular, by Corollary 3.19 we conclude that

$$\kappa_b^I \le \kappa_b, \quad \kappa_a^I \ge \kappa_a$$

for any I, where κ_b^I and κ_a^I are the constants found in Theorem 3.5 for L^I.

However, Theorem 3.5 also strongly suggests that, for any pair x, y with $x < \kappa_b < \kappa_a < y$, the restriction $L^{[x,y]}$ should be "not too different" from L itself. Specifically, we have the following result.

THEOREM 3.33. *Let L be a limit order book, and let $I = [x, y]$ with $x < \kappa_b < \kappa_a < y$. Then for all sets S, with probability 1,*

$$\limsup_{t\to\infty} \left|\mathfrak{Q}_t^b(S) - \mathfrak{Q}_t^{b,I}(S)\right| < \infty, \quad \limsup_{t\to\infty} \left|\mathfrak{Q}_t^a(S) - \mathfrak{Q}_t^{a,I}(S)\right| < \infty.$$

PROOF. This is a consequence of Theorem 3.5 and Corollary 3.15. Theorem 3.5 asserts that after a finite time there will always be a bid in L at some price in $[x, \kappa_b)$ and an ask at some price in $(\kappa_a, y]$. At that time, the states of L and L^I restricted to I differ by a finite number of orders; and Corollary 3.15 implies that this will continue to be the case at all subsequent times. □

The conjecture on the positive recurrence of the limit order book between κ_b and κ_a can be stated as follows:

CONJECTURE 3.34. For any pair x, y with $\kappa_b \le x < y \le \kappa_a$, the restriction $L^{[x,y]}$ is a Harris recurrent Markov process. Moreover, when strict inequalities $\kappa_b < x < y < \kappa_a$ hold, the restriction is positive recurrent.

If this is the case, then it makes sense to speak of the steady-state distribution of the rightmost bid and the leftmost ask for this process, and the analysis in §6 describes this distribution.

One of the things we can derive based on the conjecture is the steady-state distribution for the departing bid-ask pairs.

LEMMA 3.35. *Let $\mathcal{P}(x) = x$, and assume Conjecture 3.34. Then the steady-state distribution of the departing bid-ask pairs has density*

$$\mathbf{1}_{y \le x}(\varpi^b(x) + \varpi^a(y))$$

with respect to the Lebesgue measure dxdy.

PROOF. Departures of a bid at price x with an ask at price $y < x$ happen either when the rightmost bid is at x and an ask arrives at y (this happens at rate $\varpi^b(x)\frac{dy}{2}$) or when the leftmost ask is at y and a bid arrives at x (this happens at rate $\varpi^a(y)\frac{dx}{2}$). To obtain a normalised probability distribution, observe

$$\int_0^1 \int_y^1 (\varpi^b(x) + \varpi^a(y))dxdy = \int_0^1 \int_y^1 \varpi^b(x)dx + \int_0^1 y\varpi^a(y)dy = \mathbb{E}\varpi^b + \mathbb{E}\varpi^a.$$

Since the bid and ask distributions are symmetric, we clearly have $\mathbb{E}\varpi^b + \mathbb{E}\varpi^a = 1$, so the normalised distribution is as stated. □

For another interpretation of the restricted limit order book in terms of market orders, see §10.

8. Lyapunov function

In this section we present the proof of Theorem 3.10. The proof uses Lyapunov function techniques to show positive recurrence of the Markov chain.

We consider an ordinary limit order book with 5 equal bins,[4] each having width 1/5. We will show that the restriction of this limit order book to an interval $[\frac{1}{5} + \epsilon, \frac{4}{5} - \epsilon]$ for any $\epsilon > 0$ is positive recurrent. This implies $\kappa_b \le \frac{1}{5}$ for the ordinary limit order book. Corollary 3.25 then implies that $\kappa_b \le \frac{1}{4}$ for the strict limit order book with 4 bins, and hence $\kappa_b \le \frac{1}{4}$ for $\mathcal{P}(x) = x$. Since $\mathcal{P}(x) = x$ refines all partial orderings we consider, Corollary 3.19 implies $\kappa_b \le \frac{1}{4}$ always (and similarly, $\kappa_a \ge \frac{3}{4}$ always).

It remains to show that the restriction of the ordinary limit order book with 5 equal bins to $[\frac{1}{5} + \epsilon, \frac{4}{5} - \epsilon]$ is positive recurrent. We will refer to the bins in of the restricted limit order book as 2, 3, 4 (inheriting the numbering from the original limit order book).

[4] The number 5 isn't magical, it's just the largest number we can comfortably analyze: working with N bins would produce an $(N-2)$-dimensional Markov chain below; for $N = 4$ it turns out to be boring, and for $N = 6$ difficult to work with.

It is clearly sufficient to consider the Markov chain with state space in $\mathbb{Z}^3$ which counts the number of orders of each type in each of the three bins; the sign will be positive if the orders are bids, and negative if they are asks. We denote the state of this Markov chain

$$X_t \equiv (X_t(2), X_t(3), X_t(4)).$$

We will show that this Markov chain is positive recurrent. (This implies that the restricted limit order book has finite mean recurrence time associated with the state "empty".)

To do this, we will construct a Lyapunov function $\mathcal{L} = \mathcal{L}(X_t)$. Write

$$\Delta\mathcal{L}(t) \equiv \mathcal{L}(X_{t+1}) - \mathcal{L}(X_t), \quad \Delta X_t \equiv X_{t+1} - X_t.$$

Our Lyapunov function will have the property that $|\Delta\mathcal{L}(t)|$ is uniformly bounded, and for all sufficiently large states, $\mathbb{E}[\Delta\mathcal{L}(t)|X_t] < -\epsilon < 0$. The Foster-Lyapunov criterion asserts that existence of such a Lyapunov function implies positive recurrence of the Markov chain. (See e.g. [Bramson, 2006, Proposition 4.4], and references therein).

We now describe the construction of the Lyapunov function. $\mathcal{L}$ will have the form

$$\mathcal{L}(X) \equiv \min_F \langle X, v_F \rangle$$

for some finite set of vectors v_F which we will find below (102), (104). Therefore, level sets of $\mathcal{L}$ will be polyhedra

$$P^k \equiv \{x : \mathcal{L}(x) = k\} = kP^1$$

whose faces have v_F as their outer normals. Informally, the drift $\mathbb{E}[\Delta\mathcal{L}(t)|X_t]$ ought to be negative provided

$$\langle \mathbb{E}[\Delta X_t | X_t], v_F \rangle < 0 \tag{100}$$

for all faces F of $P^{\mathcal{L}(X_t)}$ to which the state X_t belongs[5].

We now set out to find an appropriate set v_F. Note that $\mathbb{E}[X_{t+1} - X_t|X_t]$ has only ten different values, depending on $[\alpha_t]$, $[\beta_t]$ (i.e. on the sign of the coordinates $X(2), X(3), X(4)$), namely:

(101) $\mathbb{E}[X_{t+1} - X_t|X_t] =$

$$\begin{cases} \Delta_{+++} \equiv (\frac{1}{5} - \epsilon, \frac{1}{5}, -(\frac{4}{5} - \epsilon)), & X(4) > 0 \\ \Delta_{---} \equiv (\frac{4}{5} - \epsilon, -\frac{1}{5}, -(\frac{1}{5} - \epsilon)), & X(2) < 0 \\ \Delta_{++-} \equiv (\frac{1}{5} - \epsilon, -\frac{3}{5}, \frac{2}{5} - \epsilon), & X(3) > 0,\ X(4) < 0 \\ \Delta_{+--} \equiv (-(\frac{2}{5} - \epsilon), \frac{3}{5}, -(\frac{1}{5} - \epsilon)), & X(2) > 0,\ X(3) < 0 \\ \Delta_{++0} \equiv (\frac{1}{5} - \epsilon, -\frac{3}{5}, 0), & X(3) > 0,\ X(4) = 0 \\ \Delta_{0--} \equiv (0, \frac{3}{5}, -(\frac{1}{5} - \epsilon)), & X(2) = 0,\ X(3) < 0 \\ \Delta_{+0-} \equiv (-(\frac{2}{5} - \epsilon), 0, \frac{2}{5} - \epsilon), & X(2) > 0,\ X(3) = 0,\ X(4) < 0 \\ \Delta_{+00} \equiv (-(\frac{2}{5} - \epsilon), 0, 0), & X(2) > 0,\ X(3) = X(4) = 0 \\ \Delta_{00-} \equiv (0, 0, \frac{2}{5} - \epsilon), & X(2) = X(3) = 0,\ X(4) < 0 \\ \Delta_{000} \equiv (0, 0, 0), & X(2) = X(3) = X(4) = 0 \end{cases}$$

Of these, the last one is completely irrelevant to our purposes because it only applies to a one-point set.

We will index the orthants of $\mathbb{R}^3$ and the faces between them by the signs of $X(2)$, $X(3)$, $X(4)$. Note that we only care about the closure of four orthants: $+++$, $++-$,

[5]This assumes that the only relevant directions are the faces containing X_t, which may not be the case. Instead, the binding direction for X_{t+1} could, in principle, be different from the one for X_t. However, if we start from a large state, and have small step size, this will not be a problem. We will return to this point below, after finding the relevant vectors v_F.

$+--$, and $---$. The level set $P \equiv P^1$ will be constructed by starting off with the four faces of a unit octahedron, corresponding to

$$\begin{aligned} v_{+++} &\equiv (1,1,1), & v_{++-} &\equiv (1,1,-1), \\ v_{+--} &\equiv (1,-1,-1), & v_{---} &\equiv (-1,-1,-1). \end{aligned} \tag{102}$$

This choice satisfies (100) whenever X_t lies in the interior of an orthant, but not on the boundaries between them. We will now find vectors v_{++0}, v_{0--}, v_{+0-}, v_{+00}, and v_{00-} satisfying the conditions below, which will be normal to P at points where it crosses the boundaries between orthants.

$$\langle v_{++0}, \Delta_{+++}\rangle < 0, \quad \langle v_{++0}, \Delta_{++0}\rangle < 0, \quad \langle v_{++0}, \Delta_{++-}\rangle < 0 \tag{103a}$$

$$\langle v_{0--}, \Delta_{---}\rangle < 0, \quad \langle v_{0--}, \Delta_{0--}\rangle < 0, \quad \langle v_{0--}, \Delta_{+--}\rangle < 0 \tag{103b}$$

$$\langle v_{+0-}, \Delta_{++-}\rangle < 0, \quad \langle v_{+0-}, \Delta_{+0-}\rangle < 0, \quad \langle v_{+0-}, \Delta_{+--}\rangle < 0 \tag{103c}$$

$$\begin{aligned} &\langle v_{+00}, \Delta_{+++}\rangle < 0, \quad \langle v_{+00}, \Delta_{++0}\rangle < 0, \quad \langle v_{+00}, \Delta_{++-}\rangle < 0, \\ &\langle v_{+00}, \Delta_{+00}\rangle < 0, \quad \langle v_{+00}, \Delta_{+0-}\rangle < 0, \quad \langle v_{+00}, \Delta_{+--}\rangle < 0 \end{aligned} \tag{103d}$$

$$\begin{aligned} &\langle v_{00-}, \Delta_{---}\rangle < 0, \quad \langle v_{00-}, \Delta_{0--}\rangle < 0, \quad \langle v_{00-}, \Delta_{--+}\rangle < 0, \\ &\langle v_{00-}, \Delta_{00-}\rangle < 0, \quad \langle v_{00-}, \Delta_{+0-}\rangle < 0, \quad \langle v_{00-}, \Delta_{++-}\rangle < 0 \end{aligned} \tag{103e}$$

We also require the vectors v to be *outer* normals at the relevant points of P, so the coordinates of the v's must have the same sign as the corresponding index. For example, writing $v_{++0} = (v_2, v_3, v_4)$, we must have $v_2 > 0$ and $v_4 > 0$; v_3 can have either sign.

There is a subtle point here. When we go to solve the inequalities (103), the vector v_{+0-} satisfies the following. If X_t belongs to one of the orthants adjacent to the edge given by this list of signs or lies on the edge itself (i.e. $X(2) > 0$, $X(4) < 0$, and $X(3)$ has either sign or is equal to 0), then the dot product of the one-step drift from X_t and v_{+0-} is negative. Consequently, when we go to cut up the octahedron, we must make sure that points of the resulting level set where the outer normals are given by v_{+0-} lie in one of the two orthants $++-$, $+--$, or on the boundary between them, and similarly for all the other vectors.

Provided all of the above is satisfied, we will have constructed a polyhedron P with the property that whenever X_t is on P, the vector $\mathbb{E}[X_{t+1} - X_t | X_t]$ always points strictly into P. Since $\|X_{t+1} - X_t\|$ is bounded (by 1), our construction guarantees that, starting from all sufficiently large states, (100) will hold; moreover, since we have a finite number of vectors v_F, the drift will be bounded away from 0.

Verifying that a solution to these inequalities exists for all sufficiently small $\epsilon > 0$ is a simple computation: for example, we may take

$$\begin{aligned} v_{++0} = v_{+00} &\equiv \left(\frac{4}{3}, 1, \frac{2}{3}\right), \\ v_{+0-} &\equiv \left(1, -\frac{4}{5}, -\frac{9}{5}\right), \\ v_{0--} = v_{00-} &= (-2,-3,-4). \end{aligned} \tag{104}$$

Finally, the function $\mathcal{L}$ is given by

$$\mathcal{L}(X) = \min\{\langle X, v_F\rangle\}, F \in \{+++, ++-, +--, ---, +00, +0-, 00-\}. \tag{105}$$

where the vectors v_F are given by (102) and (104).

In Figure 3.1 we show the constructed level set P. We only care about the four orthants (with boundary) which are physically possible for the limit order book, i.e. $+++$, $++-$, $+--$, and $---$; hence the resulting polyhedron is not convex.

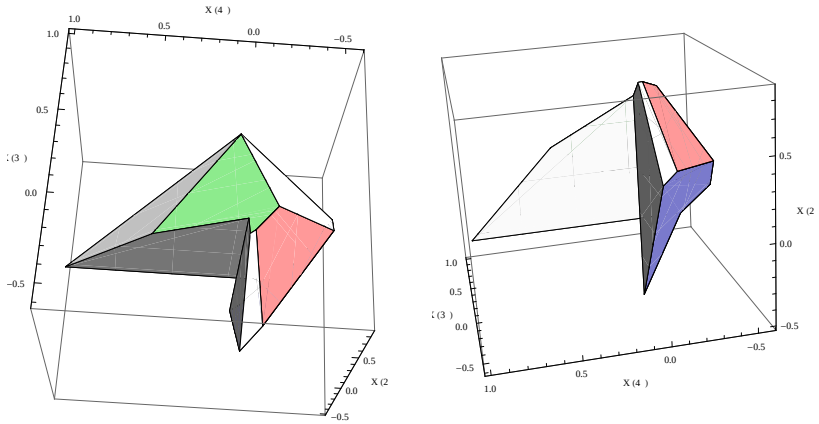

FIGURE 3.1. Two views of the level set P for the Lyapunov function (105). Red face is normal to v_{+0-}, green is normal to v_{+00}, blue is normal to v_{00-}. The remaining planes come from the octahedron $|X(2)|+|X(3)|+|X(4)| = 1$ and the coordinate planes.

The vertices and faces of this polyhedron are listed in Appendix 4.

REMARK 3.36. The method of constructing a Lyapunov function by building a polyhedral level set can in principle be carried out for finer price level functions. However, the the procedure of starting with the octahedron $\sum |x_i| = \text{const.}$ and "filing away" at the vertices and edges with additional hyperplanes that satisfy the appropriate constraints becomes substantially more difficult as the number of dimensions increases.

9. Arrival distributions

In this section we discuss the behaviour of the limit order book if the arrivals are iid, but the distribution of prices is not uniform.

We will assume that arriving orders are equally likely to be bids and asks. (If this is not satisfied, it is clear that one of the order types will have a queue going off to infinity.)

First, suppose that in a limit order book L prices of arriving bids and of arriving asks have the same distribution F, i.e.

$$\mathbb{P}(A_t \in dp \times \text{bid}) = \mathbb{P}(A_t \in dp \times \text{ask}) = dF(p).$$

The transformation

$$x \mapsto \frac{1}{2\pi}(\arctan(x) + \pi)$$

maps $\mathbb{R}$ to $(0,1)$, so (reparametrising the price if necessary) we may assume that the support of F is contained in $[0,1]$. As always, we use the convention that F is RCLL (right-continuous with left limits). Let $F^{-1} : (0,1] \to [0,1]$ be defined by

$$F^{-1}(y) = \inf\{x : F(x) \geq y\}.$$

Let the price level function of L be $\mathcal{P} : [0,1] \to [0,1]$.

Consider now a limit order book $\tilde{L}$ whose arrivals are iid uniform on $[0,1] \times \{\text{bid}, \text{ask}\}$, and apply the map $x \mapsto F^{-1}(x)$; this reproduces the arrival processes of L. Thus, if we take the price level function on $\tilde{L}$ to be $\tilde{\mathcal{P}} \equiv \mathcal{P}F^{-1}$, we will reproduce the dynamics of L in the following sense:

PROPOSITION 3.37. *Let L be a limit order book whose arrivals are iid, and*

$$\mathbb{P}(A_t \in dp \times bid) = \mathbb{P}(A_t \in dp \times ask) = dF(p)$$

for some distribution F on $[0,1]$; let $\mathcal{P}$ be the associated price level function. Let $\tilde{L}$ be a limit order book whose arrivals are iid uniform on $[0,1] \times \{bid, ask\}$, and whose price level function is $\tilde{\mathcal{P}} \equiv \mathcal{P}F^{-1}$. Then, defining for a counting measure $\mathfrak{M} = \sum m_i \delta_{x_i}$ and a function g the composition $g(\mathfrak{M}) \equiv \sum m_i \delta_{g(x_i)}$, we have equality in distribution

$$F^{-1}\tilde{\mathfrak{A}}^i_t \stackrel{d}{=} \mathfrak{A}^i_t, \quad F^{-1}\tilde{\mathfrak{Q}}^i_t \stackrel{d}{=} \mathfrak{Q}^i_t$$

for $i = a, b$ and all times $t \geq 0$.

As mentioned above, restricting the support of F to lie within $[0,1]$ does not lose generality, because we can always reparametrise the prices so that this holds.

In the case where the partial ordering of price levels is simply the total ordering on $\mathbb{R}$ ($\mathcal{P}(x) = x$) and the arrival distribution is absolutely continuous with positive density $f(x)$, instead of modifying the price level function $\mathcal{P}$, we can simply add an extra coordinate transformation. Applying this intuition to Theorem 3.5 and Theorem 3.10, we obtain the following:

COROLLARY 3.38. *Suppose that the arrival events are iid, with*

$$\mathbb{P}(A_t \in dp \times bid) = \mathbb{P}(A_t \in dp \times ask) = dF(p)$$

for an absolutely continuous distribution F with strictly positive density f on $\mathbb{R}$. Suppose further that the starting state of the limit order book is some deterministic finite state $(\mathfrak{Q}^b_0, \mathfrak{Q}^a_0)$. Then there exist two constants κ_b and κ_a, with $|\kappa_i| < \infty$, such that the following hold for any $\epsilon > 0$, with probability 1.

1. *$\mathfrak{D}^b_\infty(-\infty, \kappa_b - \epsilon] < \infty$, and $\mathfrak{D}^a_\infty[\kappa_a + \epsilon, \infty) < \infty$. That is, there are finitely many bid departures at prices $< \kappa_b - \epsilon$, and finitely many ask departures at prices $> \kappa_a + \epsilon$.*
2. *The event $\{\mathfrak{Q}^b_t[\kappa_b + \epsilon, \infty) = 0\}$ occurs infinitely often, and the event $\{\mathfrak{Q}^a_t(-\infty, \kappa_a - \epsilon] = 0\}$ occurs infinitely often.*

The constants κ_b and κ_a are given by

$$\kappa_b = F^{-1}(\kappa), \quad \kappa_a = F^{-1}(1 - \kappa)$$

where $\kappa \approx 0.2178$ is defined by Theorem 3.10.

That is, even when the distribution of the prices of the arriving orders has infinite support, there will only be departures of bids and asks from a finite interval of prices.

REMARK 3.39. Proposition 3.37 and Corollary 3.38 contain no new mathematics, but the result of Corollary 3.38 is, perhaps, surprising: no matter what the distribution, on $\mathbb{R}$, of arriving bids and asks is (provided it's the same distribution for both of them), trading will only occur on a finite interval of prices.

We can also consider the case when arrivals of bids and asks are iid, but follow *different* distributions. We will assume that the prices of arriving bids and asks have well-defined, continuously differentiable densities on $[0,1]$. (Restricting to $[0,1]$ is no loss of generality since we can always reparametrise the price.) We will consider continuous pricing, i.e. $\mathcal{P}(x) = x$.

Consider a limit order book L with some deterministic finite starting state, price level function $\mathcal{P}(x) = x$, and arrival events $(A_t)_{t \geq 0}$ which are iid. Let the distribution of A_t be given by

$$\mathbb{P}(A_t \in dp \times \text{bid}) = \frac{1}{2} dF^b(p), \quad \mathbb{P}(A_t \in dp \times \text{ask}) = \frac{1}{2} dF^a(p)$$

for some pair of probability distributions F^b, F^a on $[0,1]$ with continuously differentiable, bounded densities f^b, f^a respectively; let

$$M = \max_{i=a,b} \sup_{p\in[0,1]} f^i(p).$$

The analysis of Section 6 can be carried through more or less verbatim to derive limiting densities $\varpi^{i,f}$ $(i = a, b)$ satisfying the following equations:

$$F^a(x)\varpi^{b,f^b}(x) = f^b(x)\int_x^1 \varpi^{a,f^a}(y)dy, \quad \int_{\kappa_b}^{\kappa_a} \varpi^{b,f^b}(x)dx = 1, \tag{106a}$$

$$(1-F^b(x))\varpi^{a,f^a}(x) = f^a(x)\int_0^x \varpi^{b,f^b}(y)dy, \quad \int_{\kappa_b}^{\kappa_a} \varpi^{a,f^a}(x)dx = 1 \tag{106b}$$

with the boundary condition

$$\varpi^{b,f}(\kappa_a) = \varpi^{a,f}(\kappa_b) = 0. \tag{106c}$$

We can proceed as in Section 6 to derive ordinary differential equations for $\varpi^{b,f}$, which in particular determine the threshold values κ_b and κ_a in the case of (almost) arbitrary iid arrivals.

THEOREM 3.40. *Let L be a limit order book with some deterministic finite starting state, price level function $\mathcal{P}(x) = x$, and arrival events $(A_t)_{t\geq 0}$ which are iid. Suppose that the distribution of arrival prices of bids, respectively asks, has bounded, continuously differentiable density f^b, respectively f^a. Then the threshold values κ_b and κ_b identified in Theorem 3.5 are given as the unique values for which the system of equations* (106) *has a solution.*

If we restrict our attention to *symmetric* arrival distributions, i.e.

$$f^a(p) = f^b(1-p) \tag{107}$$

above, then (106) can be written in a particularly satisfying form. Namely, we obtain

$$\varpi^{b,f}(x)F^a(x) = f^a(1-x)\Pi^{b,f}(1-x), \quad x \in [\kappa, 1-\kappa] \tag{108}$$

where $\Pi^{b,f}(x) = \int_0^x \varpi(p)dp$. In particular, we see that the roles played by the arrival distribution and the distribution of the extreme order (highest bid or lowest ask) are in a certain sense symmetric.

If the arrival distributions of bids and asks are not the same, Theorem 3.9 no longer holds, so we no longer can conclude that $\kappa_b > 0$. Indeed, it is easy to come up with examples of densities f^b, f^a for which this is not the case. For example, taking $f^b = 2\cdot\mathbf{1}_{[1/2,1]}$ and $f^a = 2\cdot\mathbf{1}_{[0,1/2]}$, any bid-ask pair is eligible to depart, and the entire system becomes essentially a single two-sided queue, or a symmetric random walk. The inequality need not be so extreme; the analysis of restricted limit order books in §7 implies that we can take f^b to be uniform on $(\kappa, 1]$ and f^a to be uniform on $[0, 1-\kappa)$, for $\kappa = \kappa_b \approx 0.2178$ defined by Theorem 3.11.

We might like to know whether a pair of continuously differentiable, bounded distributions f^b, f^a with

$$0 < \inf_{p\in[0,1]} \frac{f^b(p)}{f^a(p)} \leq \sup_{p\in[0,1]} \frac{f^b(p)}{f^a(p)} < \infty$$

can have $\kappa_b = 0$ and $\kappa_b = 1$, i.e. distributions ϖ^b, ϖ^a which are positive on $(0,1)$. The answer is negative, as the following argument shows.

Note that we must have f^a, f^b strictly positive on $(0,1)$ (bids cannot remain waiting in a region where they do not arrive). We may therefore reparametrise so that $f^a(p) = 1$,

and f^b is some distribution that is bounded from below and above. Solving (106) for f^b, we obtain

$$f^b(x) = \frac{x}{1-x}\frac{(1-x)\pi^b(x)}{\int_x^1 \pi^a(y)dy}.$$

Take the limit as $x \to 0$. The last term converges to 1; while the first term clearly tends to 0. This contradicts the assumption that f^b is bounded away from 0.

If we remove the assumption of bounded ratios and simply require both densities to be positive on all of $(0,1)$, it is certainly possible to have $\varpi^b > 0$ on all of $(0,1)$. For example, in the symmetric case, taking the (unbounded) ask arrival density $f^a(x) = \frac{1}{2\sqrt{x}} = f^b(1-x)$ and solving the resulting (108) *does* give a density $\varpi^{b,f}$ whose support is equal to the entire interval $[0,1]$ (here, $\varpi^{b,f} = f^a$). However, the techniques used in the proof of Theorem 3.7 do not apply for unbounded distributions, so it is not clear whether $\varpi^{b,f}$ has an interpretation as the empirical distribution in this setting.

10. Market orders

In all of the above discussion we have explicitly assumed that all orders arrive as limit orders, although some of them will be executed immediately upon arrival. We can also consider a model in which there is a steady stream of *market orders* arriving at rate θ.

Definition 3.41. A *market bid* (respectively *market ask*) arriving at time t matches the lowest available waiting limit ask (respectively highest available waiting limit bid) in the system. If no asks (respectively no bids) are waiting in the system, the market bid (respectively market ask) is cancelled immediately, and disappears.

A market order is submitted by impatient traders, who want a trade to be executed immediately. We will assume that orders are not submitted when there is nobody waiting to trade. Equivalently, we may think that there is an infinite supply of waiting asks at price "$+\infty$" (very high), and an infinite supply of waiting bids at price "$-\infty$" (very low); in real-world financial markets, such liquidity might be provided by *market makers*[6].

Note that some of the "limit" orders we have considered so far are also matched immediately, so in a sense the model we have been considering already has a notion of "market orders" as orders that do not have to wait to be executed. Let us refer to limit orders which are executed immediately as *pseudo-market* orders, to distinguish them from the new stream of market orders we are introducing in this section. Pseudo-market orders have the property that when the bid price β_t is high, the number of pseudo-market asks increases; when the ask price α_t is low, the number of pseudo-market bids increases. This is a natural dynamic – if the "price" of impatience decreases, we would expect more impatient orders. What we add now is an extra stream of market orders, corresponding to the assumption that a fraction of traders are impatient irrespective of the price.

Let $A_m^b(t)$ (respectively $A_m^a(t)$) denote the number of market bids (respectively asks) that have arrived up to time t. Consider a limit order book L whose arrivals happen in discrete time, and are supported on $[0,1]$. Assume further that $A_m^b(t)$ is binomial with parameter $\theta_b t$, and $A_m^a(t)$ binomial with parameter $\theta_a t$, independently of all the other arrival and departure processes.

We can construct a process with the same distribution of waiting orders as follows. Speed up time, so that arrivals happen at times $\tau_t = \frac{t}{1+\theta_a+\theta_b}$. With probability $\frac{\theta_a}{1+\theta_a+\theta_b}$, let the incoming arrival be a market ask, which we will model as a pseudo-market ask with

[6]A *market maker* is an agent in the market who offers to buy and sell large quantities of the commodity, at low and high prices respectively. The difference between the buying and selling price pays for the risks associated with providing liquidity to the market.

price uniformly distributed over $[1, 1+\theta_a]$. With probability $\frac{\theta_b}{1+\theta_a+\theta_b}$, let the incoming arrival be a market bid, which we will model as a pseudo-market bid with price uniformly distributed over $[-\theta_b, 1]$. Finally, with the remaining probability $\frac{1}{1+\theta_a+\theta_b}$, let the incoming arrival be a limit order, which we will model as the first arrival event from the original arrival process that has not yet occurred.

Up to rescaling the price by a factor of $1+\theta_a+\theta_b$, this is the same construction as was used in the restricted limit order book in §7. Consequently, we see that restricted limit order books can be naturally interpreted as limit order books with market orders. In particular, if arrivals are iid uniform on $[0,1]$, and $\theta_a = \theta_b = \theta$, we expect qualitatively different behaviour for $\theta < \frac{\kappa}{1-2\kappa} \approx 0.386$ and for $\theta > \frac{\kappa}{1-2\kappa}$ (where $\kappa = \kappa_b$ is as defined in Theorem 3.10). Specifically, we conjecture the following:

CONJECTURE 3.42. Suppose limit order arrivals are iid uniform on $[0,1] \times \{\text{bid}, \text{ask}\}$, and suppose market bids and market asks arrive at rate θ as above, independently of the state of the system. Suppose $\theta > \frac{\kappa}{1-2\kappa}$, where $\kappa = \kappa_b$ is as defined in Theorem 3.10. Then the Markov chain describing the waiting orders in the system is positive recurrent.

For $\theta < \frac{\kappa}{1-2\kappa}$ we know that the Markov chain is *not* positive recurrent, because the corresponding restricted limit order book is restricted to an interval containing $[\kappa, 1-\kappa]$, and consequently, the number of waiting bids and waiting asks is tending to infinity (Theorem 3.33).

11. Simulation results

In this section we show a few results of simulating the limit order book with $N = 100$ equally spaced bins, and uniform arrivals. The simulations were done in Maple.

Figure 3.2 plots the distribution ϖ^b (in green) described in Proposition 3.30, along with the empirical distribution (in red) of the rightmost bid for a limit order book with $N = 100$ price bins over a sufficiently long time. The close agreement between the empirical distribution and the curve ϖ^b suggests that the rightmost bid really does have a steady-state distribution, ϖ^b.

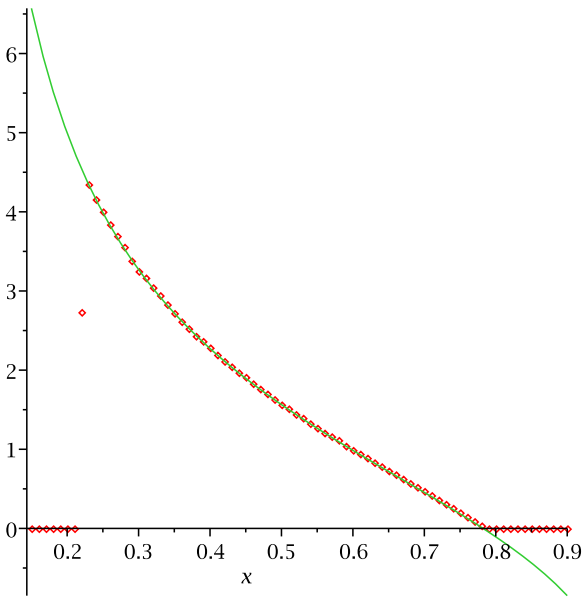

FIGURE 3.2. Distribution of the rightmost bid, together with the predicted curve.

Figure 3.3 presents the joint empirical distribution of the highest bid and the lowest ask with $N = 100$ bins. It seems plausible that the pair (β_t, α_t) has a true steady-state density on $[0, 1]^2$ (although we have not proved anything about the joint distribution even along subsequences; all our analysis was concerned with the marginal distributions). The distribution is supported on a triangle because we always have $\beta(t) < \alpha(t)$; the wide strip around the distribution corresponds to the threshold values κ_b and κ_a. Although the density appears to dip sharply as we approach the corner $(\beta_t, \alpha_t) = (\kappa_b, \kappa_a)$, it appears to be positive everywhere except possibly the corner itself. This supports the conjecture that the restriction of a limit order book to any interval $[\kappa_b + \epsilon, \kappa_a - \epsilon]$ with $\epsilon > 0$ is positive Harris recurrent.

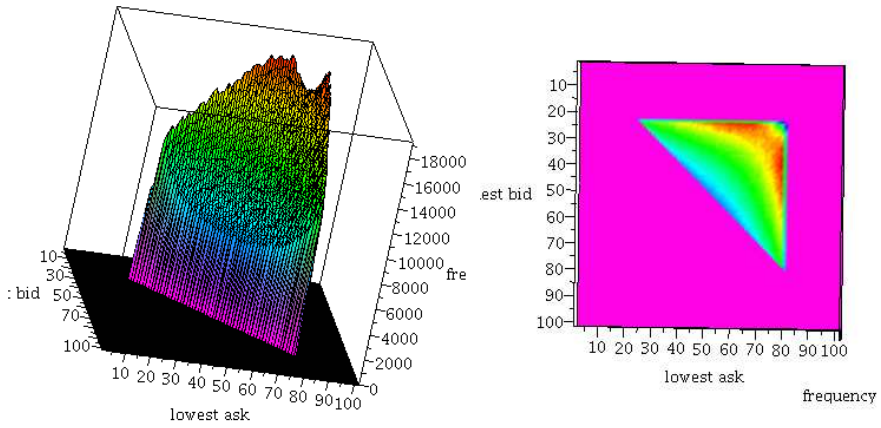

FIGURE 3.3. Joint distribution of the rightmost bid and the leftmost ask.

Bibliography

M. Armony and A. R. Ward. Blind fair routing in large-scale service systems. Submitted, 2011. http://www.stern.nyu.edu/om/faculty/armony/ArWa_10_6_11.pdf.

R. Atar, Y. Shaki, and A. Shwartz. A blind policy for equalizing cumulative idleness. *Queueing Systems*, 67(4):275–293, 2011.

F. Baccelli and T. Bonald. Window flow control in FIFO networks with cross traffic. *Queueing Systems*, 32:195–231, 1999.

F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2): 248–260, 1975.

B. Bradie. *A friendly introduction to numerical analysis*. Prentice Hall, 2006.

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

M. Bramson. Instability of FIFO queueing networks. *Annals of Applied Probability*, 4(2): 414–431, 1994.

M. Bramson, J. G. Dai, and J. M. Harrison. Positive recurrence of reflecting Brownian motion in three dimensions. *Annals of Applied Probability*, 20(2):753–783, 2010.

G. Brigham. On a congestion problem in an aircraft factory. *Journal of the Operations Research Society of America*, 3(4):412–428, 1955.

R. Caldentey, E. Kaplan, and G. Weiss. FCFS infinite bipartite matching of servers and customers. *Advances in Applied Probability*, 41:695–730, 2009.

R. Cont and A. de Larrard. Price dynamics in a Markovian limit order market. Working paper, 2010. http://ssrn.com/abstract=1735338.

M. Csörgő and L. Horváth. *Weighted approximations in probability and statistics*. Wiley, 1993.

J. G. Dai and T. Tezcan. State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research*, 36(2):271–320, 2011.

R. Delgado, F. J. López, and G. Sanz. Local conditions for the stochastic comparison of particle systems. *Advances in Applied Probability*, 36(4):1252–1277, 2004.

V. Dumas. A multiclass network with non-linear, non-convex, nonmonotonic stability conditions. *Queueing Systems*, 25:1–43, 1997.

A. El Kharroubi, A. Ben Tahar, and A. Yaacoubi. Sur la récurrence positive du mouvement Brownien réflechi dans l'orthant positif de $\mathbb{R}^n$. *Stochastics and Stochastics Reports*, 68(3-4):229–253, 2000.

A. El Kharroubi, A. Ben Tahar, and A. Yaacoubi. On the stability of the linear Skorohod problem in an orthant. *Mathematical Methods of Operations Research*, 56(2):243–258, 2002.

M. Farkas. *Dynamical Models in Biology*. Academic Press, 2001.

F. G. Foster. A unified theory for stock, storage and queue control. *Operations Research*, 10(3):121–130, 1959.

D. Gamarnik and D. Katz. Stability of Skorokhod problem is undecidable. Submitted, 2010. arXiv:1007.1694v1.

D. Gamarnik and P. Momcilovic. Steady-state analysis of a multiserver queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 40:548–577, 2008.

D. Gamarnik and A. L. Stolyar. Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime. Asymptotics of the stationary distribution. *Queueing Systems*, pages 1–27, 2012. URL `http://dx.doi.org/10.1007/s11134-012-9294-x`. To appear in print.

D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized jackson networks. *The Annals of Applied Probability*, 16:56–90, 2006.

D. K. Gode and S. Sunder. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy*, 101(1):119–137, 1993.

M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. In preparation, 2011. `arXiv:1012.0349v2`.

I. Gurvich and W. Whitt. Queue-and-Idleness-Ratio controls in many-server service systems. *Mathematics of OR*, 34(2):363–396, 2009.

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

J. M. Harrison. Brownian models of open processing networks: Canonical representation of workload. *The Annals of Applied Probability*, 10(1):75–103, 2000.

J. M. Harrison and M. J. López. Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33(4):339–368, 1999.

J. M. Harrison and V. Nguyen. The QNET method for two-moment analysis of open queueing networks. *Queueing Systems: Theory and Applications*, 6:1–32, 1990.

J. M. Harrison and R. J. Williams. Brownian models of open queueing networks with homogeneous customer populations. *Stochastics*, 22:77–115, 1987.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

J. R. Jackson. Networks of waiting lines. *Operations Research*, 5(4):518–521, 1957.

J. R. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963.

I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, second edition, 1996.

F. P. Kelly. Networks of queues. *Advances in Applied Probability*, 8:416–432, 1976.

F. P. Kelly and C. N. Laws. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing Systems*, 13:47–86, 1993.

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

J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistics Society*, 24: 383–392, 1962.

P. R. Kumar and T. I. Seidman. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control*, 35:289–298, 1990.

R. S. Liptser and A. N. Shiryaev. *Theory of Martingales*. Kluwer Academic Publishers, 1989. Translated from Russian by K. Dzjaparidze. In Russian published by Nauka 1986.

P. Lorek and R. Szekli. Strong stationary duality for M obius monotone Markov chains: Unreliable networks. *Queueing Systems*, pages –1–17, 2012. URL `http://dx.doi.org/10.1007/s11134-012-9284-z`. To appear in print.

A. Mandelbaum and A. L. Stolyar. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research*, 52:836–855, 2004.

A. Mandelbaum, W. A. Massey, and M. I. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998.

W. A. Massey. Stochastic orderings for Markov processes on partially ordered spaces. *Mathematics of Operations Research*, 12(2):350–367, 1987.

C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 1. SIAM, 2000.

O. A. Nielsen. *An Introduction to Integration and Measure Theory*. John Wiley & Sons, 1997.

G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys*, 4:193–267, 2007.

C. A. Parlour. Price dynamics in limit order markets. *Review of Financial Studies*, 11 (4):789–816, 1998.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

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

A. N. Rybko and A. L. Stolyar. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problemy Peredachi Informacii*, 28(3):3–26, 1992.

D. Shah and D. Wischik. The teleology of scheduling algorithms for switched networks under light load, critical load, and overload. Submitted, 2009. `http://www.cs.ucl.ac.uk/staff/d.wischik/Research/netsched2.pdf`.

A. L. Stolyar and T. Tezcan. Shadow-routing based control of flexible multi-server pools in overload. *Operations Research*, 59(6):1427–1444, 2011.

A. L. Stolyar and T. Tezcan. Control of systems with flexible multi-server pools: A shadow routing approach. *Queueing Systems*, 66:1–51, 2010.

A. L. Stolyar and E. Yudovina. Systems with large flexible server pools: Instability of "natural" load balancing. Submitted, 2010. `arXiv:1012.4140`.

A. L. Stolyar and E. Yudovina. Tightness of invariant distributions of a large-scale flexible service system under a priority discipline. Submitted, 2012. `arXiv:1201.2978`.

R. Talreja and W. Whitt. Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing. *Management Science*, 54(8):1513–1527, 2008.

J. A. van Mieghem. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals ofApplied Probability*, 5:809–833, 1995.

N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informacii*, 32(1):20–34, 1996.

L. M. Wein. Brownian networks with discretionary routing. *Operations Research*, 39(2): 322–340, 1991.

W. Whitt. Blocking when service is required from several facilities simultaneously. *AT&T Tehcnical Journal*, 64(8):1807–1856, 1985.

D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

I. B. Ziedins and F. P. Kelly. Limit theorems for loss networks with diverse routing. *Advances in Applied Probability*, 21:804–830, 1989.

APPENDIX A

Continuity of functions

In this appendix we summarise the various notions of continuity of functions that we use.

DEFINITION A.1. A function $f : \mathbb{R} \to \mathbb{R}$ is *Lipschitz continuous with constant* K if

$$|f(x) - f(y)| \leq |x - y|, \quad \forall x, y \in \mathbb{R}$$

The definition is identical for functions defined on a subset of $\mathbb{R}$. We may omit any explicit mention of the Lipschitz constant.

DEFINITION A.2. A function $f : \mathbb{R} \to \mathbb{R}$ is *absolutely continuous* if for all $\epsilon > 0$ there exists $\delta > 0$ such that for any finite number of disjoint intervals $(a_k, b_k)_{k=1,\ldots,n}$ with $\sum_{k=1}^{n}(b_k - a_k) < \delta$ we have $\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \epsilon$. (The definition is similar for functions defined on a subset of $\mathbb{R}$.)

Clearly, any Lipschitz function is absolutely continuous; simply take $\delta = \epsilon/K$. The converse need not be true: the function $y = x^{1/3}$ is absolutely continuous on $\mathbb{R}$, but is not Lipschitz on any interval containing the origin.

Absolutely continuous functions have the following property:

PROPOSITION A.3 (Theorem 20.8 of [Nielsen, 1997]). *Let $f : \mathbb{R} \to \mathbb{R}$ be absolutely continuous. Then the set of points x such that $f'(x)$ does not exist has Lebesgue measure 0. Moreover,*

$$f(y) - f(x) = \int_x^y f'(t)dt,$$

where the integral is the Lebesgue integral.

DEFINITION A.4. A function $f : \mathbb{R} \to \mathbb{R}$ is *right-continuous with left limits* (RCLL) if, for all x,

$$\lim_{\epsilon \to 0} f(x + \epsilon) = f(x), \quad \lim_{\epsilon \to 0} f(x - \epsilon) \text{ exists.}$$

(Here, ϵ is understood to be positive.) The definition is identical if the domain is a subset of $\mathbb{R}$. If the range is $\mathbb{R}^d$ rather than $\mathbb{R}$, the RCLL property needs to hold for each of the coordinates.

This property is also often denoted càdlàg ("continue à droite, limitée à gauche").

APPENDIX B

Another reason to restrict to a tree

In this appendix we informally describe another reason for wanting to restrict attention in Chapter 2 to a routing scheme where the allowed activities form a graph without cycles. This is based on an (apparently flawed!) intuition for how instabilities arise in networks. The examples in this section are taken from [Bramson, 2006, Chapter 3].

The earliest examples of queueing networks which exhibited unstable oscillations despite each station, on average, getting no more work than it could process were given by Kumar and Seidman [1990] and Rybko and Stolyar [1992]. In Figure B.1 we sketch the route the jobs in the Kumar-Seidman network take.

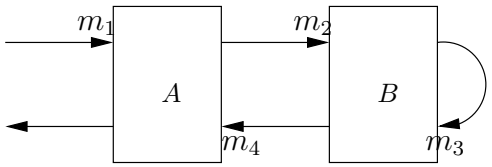

FIGURE B.1. Diagram of job flow through the Kumar-Seidman network.

We think directly about a "fluid" model. That is, each "job" has infinitesimal size, and instead we talk about the "work mass" (in a queue, or being processed by a server in a unit of time). The "work mass" arrives into the system at the top-left corner, deterministically at rate 1. Each of the rectangles A, B represents a single server, with two work stations (1 and 4 at A, 2 and 3 at B). Each job needs to visit each server twice, in the order indicated by the arrows. Service is deterministic: if the server is employed at station i, then it can process m_i^{-1} units of work per unit time. We assume that each server is individually not overloaded: $m_1^{-1} + m_4^{-1} > 1$ and $m_2^{-1} + m_3^{-1} > 1$. We will take particular values $m_2 = m_4 = \frac{2}{3}$, $m_1 = m_3 = 0$; the result holds for any values satisfying $m_2 + m_4 > 1$.

There are no routing choices here, as each job simply moves on to the next station. There is, however, a choice of scheduling, i.e. which of the stations to service. Let us suppose that the discipline is a *clearing policy*: once a server starts working on the jobs at one of its two stations, it will continue working at the same station until there is no more work in that queue. (This is a sensible policy if there are large costs associated with the server switching from one station to the other.)

Suppose the network starts empty, and work mass M enters the system at station 1. It is immediately served by server A at station 1, and passes on to station 2. At time $\frac{2}{3}M$, server B at station 2 will finish processing these initial jobs; however, an additional $\frac{2}{3}M$ jobs will enter the system in that time, be immediately processed at station 1, and join the queue at station 2, so server B continues processing the jobs at station 2. This cycle will terminate at time $2M$, when the queue at class 2 will empty out, and server B will immediately process the batch of jobs through station 3. Thus, at time $2M$ the $3M$ jobs will enter service at station 4, blocking service of new arrivals at station 1. The $3M$ jobs at station 4 will be served from time $2M$ until $4M$. During this time all jobs arriving into the system at station 1 will be forced to wait. Finally, at time $4M$ the system will again become empty, and a batch of $2M$ queued jobs will enter service at station 1. As

we iterate the process, the number of jobs queueing in the system will double at every iteration, so this network is not stable.

The reason for the instability is that jobs block the servers from doing work: if the servers were to be busy all the time, they would be able to handle the amount of work coming in, but the scheduling discipline means that it is sometimes impossible for jobs to reach the idle servers.

It seems that a salient feature of this network is the cycle present in the routing graph: jobs need to be serviced by server A, then B, then A again, and this repetition allows the system state to grow over the course of the cycle. The same feature appears in almost all of the examples in [Bramson, 2006, Chapter 3]. (An exception is the example constructed by Baccelli and Bonald [1999]; in their network, a single stream of jobs interacts with many cross-traffic streams, without ever returning to a previously visited server.) It seems intuitively plausible that restricting the routing graph to a tree would alleviate the problem, as there is "less room" to create growing excitations.

Of course, the instability examples in §5.5 demonstrate that this intuition is flawed, and that unstable, exponentially-growing excitations can occur even without any cyclic structure.

APPENDIX C

Halfin-Whitt regime

In this appendix we summarise some of the results of Halfin and Whitt [1981]. Our notation in what follows is, as much as possible, consistent with the rest of the section on call centre models, rather than with [Halfin and Whitt, 1981].

Consider a sequence of $M/M/\beta^r$ queues, indexed by a scaling parameter r[1]. The arrival process in the r^{th} system is Poisson of rate λ^r; the service in the r^{th} system is exponential with parameter μ, and there are $\beta^r \equiv r$ servers. We will be interested in the case where

$$\rho^r \equiv \frac{\lambda^r}{\mu\beta^r} = 1 - \nu r^{-1/2}$$

for some $\nu > 0$. (Here, μ and ν are constants which do not depend on r.)

REMARK C.1. Currently, our systems are indexed by the number of servers, and we carefully regulate the arrival rate into the r^{th} queue. By renumbering, we can of course equivalently think of $\lambda^r = \lambda r$ with $\beta^r = \frac{\lambda}{\mu} r + O(\sqrt{r})$.

Let $X^r(t)$ denote the number of customers in the r^{th} system (in queue and in service) at time t. Under the above scaling of arrival and service rates, the r^{th} system will be described by a positive recurrent Markov process, so $X^r(t) \to X^r$ for some limiting variable X^r. Recall that the steady-state distribution of the $M/M/\beta$ queue is

$$p_k = \mathbb{P}(X = k) = \begin{cases} \frac{1}{B}\frac{(\beta\rho)^k}{k!}, & k \le \beta \\ \frac{1}{B}\frac{\beta^\beta \rho^k}{\beta!}, & k \ge \beta \end{cases}$$

where B is a normalising constant. We can also compute the probability that an arriving customer will have to wait (note that Poisson arrivals see time averages, so it is enough to compute the steady-state probability of all β servers being occupied):

$$\mathbb{P}(X \ge \beta) = \frac{1}{B}\frac{(\beta\rho)^\beta}{\beta!(1-\rho)}$$

Note that in all of this, B will implicitly depend on r.

Cleverly expanding this for the case $\beta^r = r$ and $\rho^r = 1 - \nu r^{-1/2}$ gives:

THEOREM C.2 (Proposition 1 of [Halfin and Whitt, 1981]). *Let $\beta^r \equiv r$ and $\rho^r \equiv 1 - \nu r^{-1/2}$. Then as $r \to \infty$,*

$$\mathbb{P}(X^r \ge \beta^r) \to \alpha \equiv \left(1 + \sqrt{2\pi}\nu\Phi(\nu)\exp(\nu^2/2)\right)^{-1}.$$

Here, Φ is the cumulative distribution function of the standard normal random variable.

That is, under the $O(\sqrt{r})$ overstaffing, the probability that an arriving customer will have to wait converges to some constant between 0 and 1.

We can also consider the behaviour of the *diffusion-scaled* queue size process, $\hat{X}^r(t) = r^{-1/2}(X^r(t) - \beta^r)$. Note that this can be both positive and negative: when $\hat{X}^r(t)$ is positive, it corresponds to a positive queue, and when it is negative, it corresponds to server idleness.

[1] In β^r, r is the index rather than the exponent; we will in fact take $\beta^r = r$

THEOREM C.3 (Theorem 2 of [Halfin and Whitt, 1981]). *Let $\beta^r \equiv r$ and $\rho^r \equiv 1 - \nu r^{-1/2}$. Suppose $\hat{X}^r(0) \to \hat{X}(0)$. Then $\hat{X}^r(\cdot) \implies \hat{X}(\cdot)$ in the Skorohod space $D[0,\infty)$, where X is a diffusion process satisfying the stochastic differential equation*

$$d\hat{X}_t = m(\hat{X})dt + \sigma(\hat{X})dB_t,$$

where B is a Brownian motion (independent from all other quantities in the problem), and

$$m(x) = -\mu\nu - \mu x \mathbf{1}_{\{x \le 0\}}, \quad \sigma(x) = \sqrt{2\mu}.$$

Without proving the convergence, we argue that the drift and variance are correct. Indeed, the instantaneous drift in the r^{th} system is

$$m^r(\hat{X}^r(t)) \equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}[\hat{X}^r(t+\epsilon) - \hat{X}^r(t) | \hat{X}^r(t)]$$
$$= \begin{cases} r^{-1/2}(-r\mu + \lambda^r), & \hat{X}^r(t) > 0 \\ r^{-1/2}(-(r + \sqrt{r}\hat{X}^r(t))\mu + \lambda^r), & \hat{X}^r(t) \le 0 \end{cases}$$

which by the scaling on β^r and ρ^r equals $-\mu\nu - \mu x \mathbf{1}_{x \le 0}$. The instantaneous variance is

$$(\sigma^r)^2(\hat{X}^r(t)) \equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \mathbb{E}[(\hat{X}^r(t+\epsilon) - \hat{X}^r(t))^2 - (m^r(\hat{X}^r(t)))^2 | \hat{X}^r(t)]$$
$$= \begin{cases} r^{-1}(r\mu + \lambda^r), & \hat{X}^r(t) > 0 \\ r^{-1}((r + \sqrt{r}\hat{X}^r(t))\mu + \lambda^r), & \hat{X}^r(t) \le 0 \end{cases}$$

which converges to 2μ under the chosen scaling. Here, we use the fact that the arrival and service processes are Poisson; clearly, the variance would change for general interarrival or service times.

We now have the following limits:

$$\begin{array}{ccc} \hat{X}^r(t) & \xrightarrow[t\to\infty]{\text{positive recurrent}} & \hat{X}^r \\ \text{Theorem C.3} \Big\downarrow r\to\infty & & ? \Big\downarrow r\to\infty \\ \hat{X}(t) & \xrightarrow[t\to\infty]{?} & \hat{X} \end{array}$$

We would like to know whether there is in fact a limiting distribution $\hat{X}$ on the diffusion scale, which is simultaneously the steady-state distribution of the diffusion process $\hat{X}(t)$ (assuming it has a steady-state distribution) and the weak limit of the steady-state variables $\hat{X}^r$. Typically, we are interested in the steady-state X^r for large r, which we might attempt to approximate by the steady-state of the diffusion approximation (which is computationally "easier").

However, for this model it turns out to be relatively simple to compute the limit X^r directly:

THEOREM C.4 (Theorem 1 and Corollary 2 of [Halfin and Whitt, 1981]). *In the above set-up, $\hat{X}^r \to \hat{X}$, where $\hat{X}$ has an exponential tail with parameter ν above 0, and a normal tail below 0. More precisely, $\mathbb{P}(\hat{X} \ge 0) = \alpha$ as given by Theorem C.2, above 0 we have $\mathbb{P}(\hat{X} > x | \hat{X} \ge 0) = e^{-x\nu}$, and below 0 we have $\mathbb{P}(\hat{X} \le x | \hat{X} \le 0) = \Phi(\nu + x)/\Phi(\nu)$, where Φ is the cumulative distribution function of a standard normal variable. This is also the invariant distribution of the diffusion process $\hat{X}(\cdot)$ given by Theorem C.3.*

The argument for showing convergence $\hat{X}^r \to \hat{X}$ is similar to the argument for Theorem C.2, in that it involves explicitly approximating the steady-state distributions for an

$M/M/\beta$ queue. Note that if $\hat{X}^r \to \hat{X}$, then $\hat{X}$ must be the steady-state distribution of $\hat{X}(\cdot)$. Indeed, letting $\hat{X}^r(0) \equiv \hat{X}^r \to \hat{X} \equiv \hat{X}(0)$, we obtain at all future times $\hat{X}(t) = \hat{X}$, so $\hat{X}$ is the invariant distribution for $\hat{X}(\cdot)$.

In more general systems, it will be sufficient to prove (a) tightness of the family of distributions $\hat{X}^r$, and (b) the existence of a "nice" limiting process (e.g., solution to an SDE) $\hat{X}(\cdot)$: then any subsequential limit of $\hat{X}^r$ will be an invariant distribution of $\hat{X}(\cdot)$, which for a nice process is unique. Therefore, all convergent subsequences of $\hat{X}^r$ converge to $\hat{X}$, and under tightness, any subsequence of $\hat{X}^r$ has a convergent subsequence. We conclude that $\hat{X}^r \to \hat{X}$.

APPENDIX D

Computations

1. Computations for Example 2.32

In this section, we include the Maple code (with output) for constructing the small example of a call centre model which is unstable in underload (Example 2.32 in §2.5.5). In the multi-page formula expanding the quantity $c_2c_1 - c_0$, whose sign determines the presence of eigenvalues with positive real part in the matrix A_u, note that all but four terms are preceded by a $-$ sign (the four pluses have been replaced by $\oplus$ to make them stand out). This supports the informal Conjecture 2.46 that parameters leading to instability are rare.

```
# We show here the calculation that allows us to find
# the local instability example in underloaded.
# We consider a simple 3-customer-type,
# 4-server-type system as in the paper.
# The customer types are a, b, c;
# the server types are 1, 2, 3, 4.
# We do not divide through by B = sum_j beta_j
# in computations below; in the numerical example
# it will be equal to 1.
#
```

```
# Entries of the matrix A_u (or, rather, of
# B*A_u, as explained above):
# Diagonal entries
#}
A[aa]:=-mu[a1]*beta[1]-mu[a2]*(beta[2]+beta[3]+beta[4]):
A[bb]:=-mu[b2]*(beta[1]+beta[2])-mu[b3]*(beta[3]+beta[4]):
A[cc]:=-mu[c3]*(beta[1]+beta[2]+beta[3])-mu[c4]*beta[4]:
#}
# Off-diagonal entries
B:= beta[1]+beta[2]+beta[3]+beta[4]:
A[ab]:= A[aa]+B*mu[a2]: A[ac]:=A[aa]+B*mu[a2]:
A[ba]:= A[bb]+B*mu[b2]: A[bc]:=A[bb]+B*mu[b3]:
A[ca]:= A[cc]+B*mu[c3]: A[cb]:=A[cc]+B*mu[c3]:
#
# Matrix
A[u]:=Matrix([[A[aa], A[ab], A[ac]], [A[ba], A[bb],
  A[bc]], [A[ca], A[cb], A[cc]]]):
#
```

```
# Characteristic polynomial
with(LinearAlgebra):
charpol:=CharacteristicPolynomial(A[u], x):
#
```

```
# If the polynomial is
```

```
# x^3 - c_2*x^2 + c_1*x - c_0 then
# c_0 is the product of the roots, c_1 is the sum of
# products of pairs of roots, c_2 is the sum of the roots.
# The expression c_2*c_1 - c_0 will be
# negative if all the roots have negative real parts.
# We compute the coefficients and the expression below.
# Observe that most terms in the expression come
# with a "-" sign. However, some come with a "+" sign;
# these have the "+" circled.
# Setting the corresponding parameters to be very
# large while keeping the remaining parameters very small
# produces the desired counterexample: see below.
#
c[2]:=-coeff(charpol, x^2):
c[1]:=coeff(charpol, x):
c[0]:=-coeff(charpol, x, 0):
expr:=simplify(c[2]*c[1]-c[0]);
#
```

$$
\begin{aligned}
expr := & -2\mu_{c3}\beta_1^3\mu_{b2}\mu_{a1} - 2\mu_{c3}\beta_2^3\mu_{b2}\mu_{a2} - 2\mu_{c3}\beta_3^3\mu_{b3}\mu_{a2} - 2\mu_{c4}\beta_4^3\mu_{b3}\mu_{a2} - \\
& \mu_{c3}^2\beta_1^2\beta_4\mu_{b2} - 3\mu_{c3}^2\beta_1^2\mu_{b2}\beta_2 - \mu_{c3}^2\beta_1^2\mu_{b3}\beta_3 - \mu_{c3}^2\beta_1^2\mu_{a2}\beta_2 - \mu_{c3}^2\beta_1^2\mu_{a2}\beta_3 - \\
& 2\mu_{c3}^2\beta_1^2\beta_2\mu_{a1} - 2\mu_{c3}^2\beta_1^2\beta_3\mu_{b2} - 2\mu_{c3}^2\beta_1^2\beta_3\mu_{a1} - 2\mu_{c3}^2\beta_1\beta_3^2\mu_{a2} - 2\mu_{c3}^2\beta_1\beta_3^2\mu_{b3} - \\
& 2\mu_{c3}^2\beta_1\beta_2^2\mu_{a2} - 3\mu_{c3}^2\beta_1\beta_2^2\mu_{b2} - \mu_{c3}^2\beta_1^2\beta_4\mu_{a1} - \mu_{c3}^2\beta_2^2\beta_4\mu_{b2} - \mu_{c3}^2\beta_2^2\mu_{b3}\beta_3 - \\
& \mu_{c3}^2\beta_2^2\mu_{a1}\beta_1 - 3\mu_{c3}^2\beta_2^2\mu_{a2}\beta_3 - \mu_{c3}^2\beta_2^2\mu_{a2}\beta_4 - 2\mu_{c3}^2\beta_2^2\beta_3\mu_{b2} - 3\mu_{c3}^2\beta_2\beta_3^2\mu_{a2} - \\
& 2\mu_{c3}^2\beta_2\beta_3^2\mu_{b3} - \mu_{c3}^2\beta_3^2\mu_{b2}\beta_1 - \mu_{c3}^2\beta_3^2\mu_{b2}\beta_2 - \mu_{c3}^2\beta_3^2\mu_{b3}\beta_4 - \mu_{c3}^2\beta_3^2\mu_{a1}\beta_1 - \\
& \mu_{c3}^2\beta_3^2\mu_{a2}\beta_4 - \mu_{a2}^2\beta_2^3\mu_{c3} - \mu_{a1}^2\beta_1^3\mu_{b2} - 3\mu_{a2}^2\beta_2\mu_{c3}\beta_3^2 - 2\mu_{a2}^2\beta_2\mu_{b3}\beta_3^2 - \\
& \mu_{a2}^2\beta_2^2\mu_{b2}\beta_1 - 2\mu_{a2}^2\beta_2^2\mu_{b2}\beta_3 - 2\mu_{a2}^2\beta_2^2\mu_{b2}\beta_4 - \mu_{a2}^2\beta_2^2\mu_{b3}\beta_3 - \mu_{a2}^2\beta_2^2\mu_{b3}\beta_4 - \\
& 2\mu_{a2}^2\beta_2\mu_{b3}\beta_4^2 - \mu_{a2}^2\beta_3^2\mu_{c4}\beta_4 - \mu_{a2}^2\beta_3^2\mu_{c3}\beta_1 - 2\mu_{a2}^2\beta_3^2\mu_{c3}\beta_4 - 2\mu_{a2}^2\beta_3\mu_{c4}\beta_4^2 - \\
& \mu_{a2}^2\beta_3^2\beta_1\mu_{b3} - \mu_{a2}^2\beta_3^2\mu_{b2}\beta_2 - 3\mu_{a2}^2\beta_3^2\mu_{b3}\beta_4 - 3\mu_{a2}^2\beta_3\mu_{b3}\beta_4^2 - \mu_{a2}^2\beta_4^2\mu_{c3}\beta_2 - \\
& \mu_{a2}^2\beta_4^2\mu_{c3}\beta_3 - \mu_{c4}\beta_4\mu_{a2}^2\beta_3\beta_1 - \mu_{c4}\beta_4^2\mu_{a2}\mu_{a1}\beta_1 - \mu_{c4}\beta_4\mu_{b2}\beta_1^2\mu_{a1} - \\
& \mu_{c4}\beta_4\mu_{a1}\beta_1^2\mu_{b3} - \mu_{c4}\beta_4\mu_{a2}\beta_1^2\mu_{b2} \oplus \mu_{c4}\beta_4\mu_{a2}\beta_1^2\mu_{b3} - \mu_{c3}\beta_4\mu_{a1}\beta_1^2\mu_{a2} - \\
& \mu_{c3}\beta_4\mu_{a2}^2\beta_2\beta_1 - \mu_{c3}\beta_4\mu_{a2}^2\beta_3\beta_1 - \mu_{c3}\beta_4^2\mu_{a2}\mu_{a1}\beta_1 - \mu_{c3}\beta_4\mu_{b2}\beta_1^2\mu_{a1} - \\
& \mu_{c3}\beta_4\mu_{a1}\beta_1^2\mu_{b3} - \mu_{c3}\beta_4\mu_{a2}\beta_1^2\mu_{b2} - \mu_{c3}\beta_4\mu_{a2}\beta_1^2\mu_{b3} - \mu_{c4}\beta_4^2\beta_1\mu_{b3}\mu_{a1} - \\
& \mu_{c4}\beta_4^2\beta_1\mu_{b3}\mu_{a2} - \mu_{c4}\beta_4^2\beta_1\mu_{b2}\mu_{a1} - \mu_{c4}\beta_4^2\beta_1\mu_{b2}\mu_{a2} - \mu_{c4}\beta_4\mu_{b2}\beta_1^2\mu_{b3} - \\
& \mu_{c4}\beta_4\mu_{b2}\beta_2^2\mu_{b3} - \mu_{c4}\beta_4\mu_{b3}^2\beta_3\beta_1 - \mu_{c4}\beta_4\mu_{b3}^2\beta_3\beta_2 - \mu_{c4}\beta_4^2\mu_{b3}\mu_{b2}\beta_1 - \\
& \mu_{c4}\beta_4^2\mu_{b3}\mu_{b2}\beta_2 - \mu_{c3}\beta_4^2\beta_1\mu_{b3}\mu_{a1} - \mu_{c3}\beta_4^2\beta_1\mu_{b3}\mu_{a2} \oplus \mu_{c3}\beta_4^2\beta_1\mu_{b2}\mu_{a1} - \\
& \mu_{c3}\beta_4^2\beta_1\mu_{b2}\mu_{a2} - 2\mu_{c3}\beta_4\mu_{b2}^2\beta_1\beta_2 - \mu_{c3}\beta_4\mu_{b2}\beta_1^2\mu_{b3} - \mu_{c3}\beta_4\mu_{b2}\beta_2^2\mu_{b3} - \\
& \mu_{c3}\beta_4\mu_{b3}^2\beta_3\beta_1 - \mu_{c3}\beta_4\mu_{b3}^2\beta_3\beta_2 - \mu_{c3}\beta_4^2\mu_{b3}\mu_{b2}\beta_1 - \mu_{c3}\beta_4^2\mu_{b3}\mu_{b2}\beta_2 - \\
& 2\mu_{c4}\beta_4\mu_{b2}\beta_2^2\mu_{a2} - 2\mu_{c4}\beta_4^2\mu_{b2}\beta_2\mu_{a2} - 2\mu_{c4}\beta_4^2\mu_{b3}\beta_2\mu_{a2} - 2\mu_{c3}\beta_4\mu_{b2}\beta_2^2\mu_{a2} - \\
& 2\mu_{c3}\beta_4\mu_{b3}\beta_2^2\mu_{a2} - 2\mu_{c3}\beta_4^2\mu_{b3}\beta_2\mu_{a2} - 2\mu_{c3}\beta_1^2\mu_{b2}\beta_3\mu_{a1} - 2\mu_{c3}\beta_1^2\mu_{b2}\mu_{a2}\beta_2 - \\
& 4\mu_{c3}\beta_1^2\mu_{b2}\beta_2\mu_{a1} - 4\mu_{c3}\beta_1\mu_{b2}\beta_2^2\mu_{a2} - 2\mu_{c3}\beta_1\mu_{b3}\beta_3^2\mu_{a2} - 2\mu_{c3}\beta_2^2\mu_{b2}\mu_{a1}\beta_1 - \\
& 4\mu_{c3}\beta_2^2\mu_{b2}\mu_{a2}\beta_3 - 2\mu_{c3}\beta_2^2\mu_{b3}\beta_3\mu_{a2} - 4\mu_{c3}\beta_2\mu_{b3}\beta_3^2\mu_{a2} - 2\mu_{c3}\beta_3^2\mu_{b2}\beta_2\mu_{a2} - \\
& 4\mu_{c3}\beta_3^2\mu_{b3}\mu_{a2}\beta_4 - 2\mu_{c3}\beta_3\mu_{b3}\beta_4^2\mu_{a2} - 2\mu_{c4}\beta_4\mu_{b3}\beta_3^2\mu_{a2} - 4\mu_{c4}\beta_4^2\mu_{b3}\beta_3\mu_{a2} - \\
& 2\mu_{c3}^2\beta_1\beta_2\mu_{b3}\beta_3 - 4\mu_{c3}^2\beta_1\beta_2\mu_{a2}\beta_3 - \mu_{c3}^2\beta_1\beta_2\mu_{a2}\beta_4 - 4\mu_{c3}^2\beta_1\beta_3\mu_{b2}\beta_2 - \\
& \mu_{c3}^2\beta_1\beta_3\mu_{b3}\beta_4 - \mu_{c3}^2\beta_1\beta_3\mu_{a2}\beta_4 - \mu_{c3}\beta_1\mu_{c4}\beta_4^2\mu_{a2} - \mu_{c3}\beta_1\mu_{c4}\beta_4^2\mu_{b3} - \\
& \mu_{c3}\beta_1^2\mu_{c4}\beta_4\mu_{a2} - \mu_{c3}\beta_1^2\mu_{c4}\beta_4\mu_{b3} - 2\mu_{c3}\beta_2^2\mu_{c4}\beta_4\mu_{a2} - \mu_{c3}\beta_2^2\mu_{c4}\beta_4\mu_{b3} - \\
& \mu_{c3}^2\beta_2\beta_3\mu_{b3}\beta_4 - 2\mu_{c3}^2\beta_2\beta_3\mu_{a1}\beta_1 - 2\mu_{c3}^2\beta_2\beta_3\mu_{a2}\beta_4 - 2\mu_{c3}\beta_2\mu_{c4}\beta_4^2\mu_{a2} - \\
& \mu_{c3}\beta_2\mu_{c4}\beta_4^2\mu_{b3} - \mu_{c4}\beta_4^2\mu_{a2}^2\beta_1 - \mu_{c3}\beta_4\mu_{a1}^2\beta_1^2 - \mu_{c4}\beta_4^2\mu_{b3}^2\beta_1 - \mu_{c4}\beta_4^2\mu_{b3}^2\beta_2 - \\
& \mu_{c3}\beta_4\mu_{b2}^2\beta_1^2 - \mu_{c3}\beta_4\mu_{b2}^2\beta_2^2 - \mu_{c4}^2\beta_4^2\mu_{a2}\beta_2 - \mu_{c4}^2\beta_4^2\mu_{a2}\beta_3 - \mu_{c4}^2\beta_4^2\mu_{b3}\beta_2 - \\
& \mu_{c4}^2\beta_4^2\mu_{b3}\beta_3 - \mu_{c4}^2\beta_4^2\mu_{a2}\beta_1 - \mu_{c4}^2\beta_4^2\mu_{b3}\beta_1 - 3\mu_{b2}^2\beta_1^2\mu_{c3}\beta_2 - \mu_{b2}^2\beta_1^2\mu_{c3}\beta_3 - \\
& 3\mu_{b2}^2\beta_1\mu_{c3}\beta_2^2 - 2\mu_{b2}^2\beta_1\beta_2^2\mu_{a2} - \mu_{b2}^2\beta_1^2\beta_3\mu_{a1} - \mu_{b2}^2\beta_1^2\beta_4\mu_{a1} - \mu_{b2}^2\beta_1^2\mu_{a2}\beta_2 -
\end{aligned}
$$

$$2\mu_{b2}^2\beta_1^2\beta_2\mu_{a1} - \mu_{b2}^2\beta_2^2\mu_{c3}\beta_3 - \mu_{b2}^2\beta_2^2\mu_{a1}\beta_1 - \mu_{b2}^2\beta_2^2\mu_{a2}\beta_3 - \mu_{b2}^2\beta_2^2\mu_{a2}\beta_4 -$$
$$\mu_{b3}^2\beta_3^2\mu_{c3}\beta_1 - \mu_{a2}^2\beta_4^3\mu_{c4} - \mu_{a2}^2\beta_3^3\mu_{b3} - \mu_{a2}^2\beta_3^3\mu_{c3} - \mu_{a2}^2\beta_2^3\mu_{b2} - \mu_{a2}^2\beta_4^3\mu_{b3} -$$
$$\mu_{a1}^2\beta_1^3\mu_{c3} - \mu_{b3}^2\beta_4^3\mu_{a2} - \mu_{b3}^2\beta_4^3\mu_{c4} - \mu_{a2}^2\beta_4^2\beta_1\mu_{b3} - \mu_{a2}^2\beta_4^2\mu_{b2}\beta_2 -$$
$$\mu_{c4}\beta_4\mu_{a2}\beta_3\mu_{a1}\beta_1 - \mu_{c4}\beta_4\mu_{b2}\beta_2\mu_{a1}\beta_1 - \mu_{c4}\beta_4\mu_{a1}\beta_1\mu_{b3}\beta_2 -$$
$$3\mu_{c4}\beta_4\mu_{b2}\beta_1\mu_{a2}\beta_2 \oplus \mu_{c4}\beta_4\mu_{a2}\beta_1\mu_{b3}\beta_2 - 3\mu_{c3}\beta_4\mu_{a2}\beta_2\mu_{a1}\beta_1 -$$
$$3\mu_{c3}\beta_4\mu_{a2}\beta_3\mu_{a1}\beta_1 - \mu_{c3}\beta_4\mu_{b2}\beta_2\mu_{a1}\beta_1 - \mu_{c3}\beta_4\mu_{a1}\beta_1\mu_{b3}\beta_2 -$$
$$3\mu_{c3}\beta_4\mu_{b2}\beta_1\mu_{a2}\beta_2 - 3\mu_{c3}\beta_4\mu_{a2}\beta_1\mu_{b3}\beta_2 - \mu_{c4}\beta_4\beta_1\mu_{b3}\beta_3\mu_{a1} -$$
$$\mu_{c4}\beta_4\beta_1\mu_{b3}\beta_3\mu_{a2} - \mu_{c4}\beta_4\beta_1\mu_{b2}\beta_3\mu_{a1} - \mu_{c4}\beta_4\beta_1\mu_{b2}\beta_3\mu_{a2} -$$
$$2\mu_{c4}\beta_4\mu_{b2}\beta_1\mu_{b3}\beta_2 - \mu_{c4}\beta_4\mu_{b3}\beta_3\mu_{b2}\beta_1 - \mu_{c4}\beta_4\mu_{b3}\beta_3\mu_{b2}\beta_2 -$$
$$3\mu_{c3}\beta_4\beta_1\mu_{b3}\beta_3\mu_{a1} - 3\mu_{c3}\beta_4\beta_1\mu_{b3}\beta_3\mu_{a2} \oplus \mu_{c3}\beta_4\beta_1\mu_{b2}\beta_3\mu_{a1} -$$
$$3\mu_{c3}\beta_4\beta_1\mu_{b2}\beta_3\mu_{a2} - 2\mu_{c3}\beta_4\mu_{b2}\beta_1\mu_{b3}\beta_2 - 3\mu_{c3}\beta_4\mu_{b3}\beta_3\mu_{b2}\beta_1 -$$
$$3\mu_{c3}\beta_4\mu_{b3}\beta_3\mu_{b2}\beta_2 - 2\mu_{c4}\beta_4\mu_{b2}\beta_2\mu_{a2}\beta_3 - 2\mu_{c4}\beta_4\mu_{b3}\beta_2\mu_{a2}\beta_3 -$$
$$2\mu_{c3}\beta_4\mu_{b2}\beta_2\mu_{a2}\beta_3 - 6\mu_{c3}\beta_4\mu_{b3}\beta_2\mu_{a2}\beta_3 - 6\mu_{c3}\beta_1\mu_{b2}\beta_2\mu_{a2}\beta_3 -$$
$$2\mu_{c3}\beta_1\mu_{b3}\beta_3\mu_{a2}\beta_2 - \mu_{b3}^2\beta_3^2\mu_{c3}\beta_2 - 2\mu_{b3}^2\beta_3^2\mu_{c3}\beta_4 - \mu_{b3}^2\beta_3^2\mu_{c4}\beta_4 - 2\mu_{b3}^2\beta_3\mu_{c4}\beta_4^2 -$$
$$\mu_{b3}^2\beta_3^2\beta_1\mu_{a2} - \mu_{b3}^2\beta_3^2\mu_{a2}\beta_2 - 3\mu_{b3}^2\beta_3^2\mu_{a2}\beta_4 - 3\mu_{b3}^2\beta_3\beta_4^2\mu_{a2} - \mu_{b3}^2\beta_4^2\mu_{c3}\beta_3 -$$
$$\mu_{b3}^2\beta_4^2\beta_1\mu_{a2} - \mu_{b3}^2\beta_4^2\mu_{a2}\beta_2 - \mu_{a1}^2\beta_1^2\mu_{c3}\beta_2 - \mu_{a1}^2\beta_1^2\mu_{c3}\beta_3 - \mu_{a1}^2\beta_1^2\mu_{b2}\beta_3 -$$
$$\mu_{a1}^2\beta_1^2\mu_{b2}\beta_4 - \mu_{a1}^2\beta_1^2\mu_{b2}\beta_2 - \mu_{a2}^2\beta_2^2\mu_{c4}\beta_4 - \mu_{a2}^2\beta_2^2\mu_{c3}\beta_1 - 3\mu_{a2}^2\beta_2^2\mu_{c3}\beta_3 -$$
$$2\mu_{a2}^2\beta_2^2\mu_{c3}\beta_4 - 2\mu_{a2}^2\beta_2\mu_{c4}\beta_4^2 - \mu_{b2}^2\beta_2^3\mu_{a2} - \mu_{b2}^2\beta_2^3\mu_{c3} - \mu_{b2}^2\beta_1^3\mu_{a1} -$$
$$\mu_{b3}^2\beta_3^3\mu_{a2} - \mu_{b3}^2\beta_3^3\mu_{c3} - \mu_{c4}^2\beta_4^3\mu_{b3} - \mu_{c4}^2\beta_4^3\mu_{a2} - \mu_{b2}^2\beta_1^3\mu_{c3} - \mu_{c3}^2\beta_1^3\mu_{b2} -$$
$$\mu_{c3}^2\beta_3^3\mu_{b3} - \mu_{c3}^2\beta_3^3\mu_{a2} - \mu_{c3}^2\beta_2^3\mu_{b2} - \mu_{c3}^2\beta_2^3\mu_{a2} - \mu_{c3}^2\beta_1^3\mu_{a1} - \mu_{c4}\beta_4\mu_{a2}^2\beta_2\beta_1 -$$
$$\mu_{c4}\beta_4\mu_{a1}\beta_1^2\mu_{a2} - \mu_{c4}\beta_4\mu_{a2}\beta_2\mu_{a1}\beta_1 - 2\mu_{b2}\beta_2^2\mu_{c3}\mu_{b3}\beta_3 - 2\mu_{b2}\beta_2\mu_{c3}\beta_3^2\mu_{b3} -$$
$$2\mu_{b2}\beta_2\mu_{b3}\beta_3^2\mu_{a2} - \mu_{b2}^2\beta_2\beta_1\beta_3\mu_{a1} - \mu_{b2}^2\beta_2\beta_1\beta_4\mu_{a1} - 2\mu_{b2}\beta_2^2\mu_{b3}\beta_3\mu_{a2} -$$
$$2\mu_{b2}\beta_2^2\mu_{b3}\beta_4\mu_{a2} - 2\mu_{b2}\beta_2\mu_{b3}\beta_4^2\mu_{a2} - 2\mu_{b3}\beta_3^2\mu_{c3}\mu_{a1}\beta_1 - 2\mu_{b3}\beta_3\mu_{c3}\beta_1^2\mu_{a1} -$$
$$\mu_{b3}\beta_3\mu_{b2}\beta_1^2\mu_{a1} - 2\mu_{b3}^2\beta_3\beta_1\beta_4\mu_{a2} - \mu_{b3}\beta_3^2\beta_1\mu_{b2}\mu_{a1} - 2\mu_{b3}^2\beta_3\beta_4\mu_{a2}\beta_2 -$$
$$\mu_{b3}\beta_4\mu_{b2}\beta_1^2\mu_{a1} - \mu_{b3}\beta_4^2\beta_1\mu_{b2}\mu_{a1} - 2\mu_{a1}\beta_1^2\mu_{c3}\mu_{a2}\beta_2 - 2\mu_{a1}\beta_1^2\mu_{c3}\mu_{a2}\beta_3 -$$
$$2\mu_{a1}\beta_1\mu_{c3}\beta_3^2\mu_{a2} - 2\mu_{a1}\beta_1\mu_{c3}\beta_2^2\mu_{a2} - \mu_{a1}\beta_1\mu_{b3}\beta_3^2\mu_{a2} - 2\mu_{a1}\beta_1\mu_{b2}\beta_2^2\mu_{a2} -$$
$$\mu_{a1}\beta_1^2\mu_{b3}\beta_3\mu_{a2} - \mu_{a1}\beta_1^2\mu_{b3}\beta_4\mu_{a2} - 2\mu_{a1}\beta_1^2\mu_{b2}\mu_{a2}\beta_2 - \mu_{a1}\beta_1\mu_{b3}\beta_4^2\mu_{a2} -$$
$$2\mu_{a2}^2\beta_2\mu_{c4}\beta_4\beta_3 - 2\mu_{a2}^2\beta_2\mu_{c3}\beta_1\beta_3 - 4\mu_{a2}^2\beta_2\mu_{c3}\beta_3\beta_4 - \mu_{a2}^2\beta_2\beta_1\mu_{b3}\beta_3 -$$
$$\mu_{a2}^2\beta_2\beta_1\mu_{b3}\beta_4 - 4\mu_{a2}^2\beta_2\mu_{b3}\beta_3\beta_4 - \mu_{a2}\beta_3\mu_{b2}\beta_1^2\mu_{a1} - 2\mu_{a2}^2\beta_3\beta_1\mu_{b3}\beta_4 -$$
$$\mu_{a2}\beta_3^2\beta_1\mu_{b2}\mu_{a1} - \mu_{a2}^2\beta_3\mu_{b2}\beta_1\beta_2 - 2\mu_{a2}^2\beta_3\mu_{b2}\beta_2\beta_4 - \mu_{a2}\beta_4\mu_{b2}\beta_1^2\mu_{a1} -$$
$$\mu_{a2}\beta_4^2\beta_1\mu_{b2}\mu_{a1} - \mu_{a2}^2\beta_4\mu_{b2}\beta_1\beta_2 - \mu_{c3}^2\beta_2\beta_4\mu_{a1}\beta_1 - 2\mu_{c3}\beta_3^2\mu_{c4}\beta_4\mu_{a2} -$$
$$\mu_{c3}^2\beta_3\beta_4\mu_{b2}\beta_1 - \mu_{c3}^2\beta_3\beta_4\mu_{b2}\beta_2 - 2\mu_{c3}\beta_3^2\mu_{c4}\beta_4\mu_{b3} - 2\mu_{c3}\beta_3\mu_{c4}\beta_4^2\mu_{a2} -$$
$$2\mu_{c3}\beta_3\mu_{c4}\beta_4^2\mu_{b3} - \mu_{c3}^2\beta_3\beta_4\mu_{a1}\beta_1 - \mu_{c4}\beta_4^2\mu_{c3}\mu_{b2}\beta_1 - \mu_{c4}\beta_4^2\mu_{c3}\mu_{b2}\beta_2 -$$
$$\mu_{c4}\beta_4\mu_{c3}\beta_2^2\mu_{b2} - \mu_{c4}\beta_4\mu_{c3}\beta_1^2\mu_{a1} - \mu_{c4}\beta_4\mu_{c3}\beta_1^2\mu_{b2} - \mu_{c4}\beta_4^2\mu_{c3}\mu_{a1}\beta_1 -$$
$$2\mu_{b2}\beta_1^2\mu_{c3}\mu_{b3}\beta_3 - 2\mu_{b2}\beta_1^2\mu_{c3}\mu_{a2}\beta_3 - 2\mu_{b2}^2\beta_1\mu_{c3}\beta_3\beta_2 - 2\mu_{b2}\beta_1\mu_{c3}\beta_3^2\mu_{a2} -$$
$$2\mu_{b2}\beta_1\mu_{c3}\beta_3^2\mu_{b3} - \mu_{b2}\beta_1\mu_{b3}\beta_3^2\mu_{a2} - \mu_{b2}\beta_1^2\mu_{b3}\beta_3\mu_{a2} - \mu_{b2}\beta_1^2\mu_{b3}\beta_4\mu_{a2} -$$
$$\mu_{b2}^2\beta_1\beta_2\mu_{a2}\beta_3 - \mu_{b2}^2\beta_1\beta_2\mu_{a2}\beta_4 - \mu_{b2}\beta_1\mu_{b3}\beta_4^2\mu_{a2} - 2\mu_{c3}\beta_2\beta_1\mu_{b2}\beta_3\mu_{a1} -$$
$$2\mu_{c3}^2\beta_1\beta_4\mu_{b2}\beta_2 - 3\mu_{c3}\beta_1\mu_{c4}\beta_4\mu_{a2}\beta_2 - 3\mu_{c3}\beta_1\mu_{c4}\beta_4\mu_{a2}\beta_3 -$$
$$2\mu_{c3}\beta_1\mu_{c4}\beta_4\mu_{b3}\beta_2 - 3\mu_{c3}\beta_1\mu_{c4}\beta_4\mu_{b3}\beta_3 - 4\mu_{c3}\beta_2\mu_{c4}\beta_4\mu_{a2}\beta_3 -$$
$$3\mu_{c3}\beta_2\mu_{c4}\beta_4\mu_{b3}\beta_3 - 2\mu_{c4}\beta_4\mu_{c3}\beta_1\mu_{b2}\beta_2 - \mu_{c4}\beta_4\mu_{c3}\beta_2\mu_{a1}\beta_1 -$$
$$\mu_{c4}\beta_4\mu_{c3}\beta_3\mu_{b2}\beta_1 - \mu_{c4}\beta_4\mu_{c3}\beta_3\mu_{b2}\beta_2 - \mu_{c4}\beta_4\mu_{c3}\beta_3\mu_{a1}\beta_1 - 4\mu_{b2}\beta_1\mu_{c3}\beta_2\mu_{b3}\beta_3 -$$
$$3\mu_{b2}\beta_1\mu_{b3}\beta_3\mu_{a2}\beta_2 - 2\mu_{b2}\beta_1\mu_{b3}\beta_3\mu_{a2}\beta_4 - 3\mu_{b2}\beta_1\mu_{b3}\beta_4\mu_{a2}\beta_2 -$$
$$4\mu_{b2}\beta_2\mu_{b3}\beta_3\mu_{a2}\beta_4 - 2\mu_{b3}\beta_3\mu_{c3}\beta_2\mu_{a1}\beta_1 - 2\mu_{b3}\beta_3\beta_1\mu_{b2}\beta_4\mu_{a1} -$$
$$\mu_{b3}\beta_3\mu_{b2}\beta_2\mu_{a1}\beta_1 - \mu_{b3}\beta_4\mu_{b2}\beta_2\mu_{a1}\beta_1 - 4\mu_{a1}\beta_1\mu_{c3}\beta_2\mu_{a2}\beta_3 -$$
$$3\mu_{a1}\beta_1\mu_{b2}\beta_2\mu_{a2}\beta_3 - 3\mu_{a1}\beta_1\mu_{b2}\beta_2\mu_{a2}\beta_4 - \mu_{a1}\beta_1\mu_{b3}\beta_3\mu_{a2}\beta_2 -$$
$$2\mu_{a1}\beta_1\mu_{b3}\beta_3\mu_{a2}\beta_4 - \mu_{a1}\beta_1\mu_{b3}\beta_4\mu_{a2}\beta_2 - 2\mu_{a2}\beta_3\beta_1\mu_{b2}\beta_4\mu_{a1}$$

```
#
# Numerical values and the counterexample
#
```

```
# We find a positive term in the above expression:
```

```
coeff(coeff(coeff(coeff(expr,
  mu[a2]),mu[b3]),mu[c4]),beta[1]^2);
#
```

$$\beta_4$$

```
# We set beta_1, mu_a2, mu_b3, mu_c4 to be large
# and the rest of the parameters to be small.
# Setting beta_4 large would not be productive,
# because there may well be (and, in fact, are)
# positive terms depending on beta_4 through
# beta_4^2 or even beta_4^3.
#
eval(expr, {mu[a1] = 1, mu[a2] = 100, mu[b2] = 1,
  mu[b3] = 100, mu[c3] = 1, mu[c4] = 100,
  beta[1] = .97, beta[2] = 0.1e-1, beta[3] = 0.1e-1,
  beta[4] = 0.1e-1});
#
```

$$6464.105200$$

```
# We compute the matrix, and check that it in fact
# has a pair of eigenvalues with positive real part
#
A:=eval(A[u], {mu[a1] = 1, mu[a2] = 100, mu[b2] = 1,
  mu[b3] = 100, mu[c3] = 1, mu[c4] = 100, beta[1] = .97,
  beta[2] = 0.1e-1, beta[3] = 0.1e-1, beta[4] = 0.1e-1});
```

$$A := \begin{bmatrix} -3.97 & 96.03 & 96.03 \\ -1.98 & -2.98 & 97.02 \\ -0.99 & -0.99 & -1.9 \end{bmatrix}$$

```
Eigenvalues(A);
```

$$\begin{bmatrix} 4.45477075946149625 + 23.36891629359886 84i \\ 4.45477075946149625 - 23.36891629i \\ -17.8495415189230116 + 0.0i \end{bmatrix}$$

2. Computations for Example 2.36

In this section, we include the Maple code (with output) for demonstrating that the 21-customer-class call centre model in Example 2.36 (§2.5.5) is unstable in underload.

```
# We show that the 21-customer-type example
# is really an example of local instability.
#
#   a   b   c   d   e   f   g   h   i   j   k   l  ...
#  / \ / \ / \ / \ / \ / \ / \ / \ / \ / \ / \ / \ /
# 1   2   3   4   5   6   7   8   9  10  11  12  13
#
# ... m   n   o   p   q   r   s   t   u
#  \ / \ / \ / \ / \ / \ / \ / \ / \ / \
#  13  14  15  16  17  18  19  20  21  22
#
# The service rates are 1 to the left always,
# 1/3 to the right on the leftmost 12 edges (a2 ... l13),
# 3 to the right on the rightmost 9 edges (m14 ... u22).
#
```

```
# We construct A_u:
#
betas:=Vector(22,1): B:=22:
#
mu:=Matrix(21,22,0):
for i from 1 to 21 do mu[i,i]:=1: od:
#      service rate to the left is 1
for i from 1 to 12 do mu[i,i+1]:=1/3: od:
#      service rate to the right is 1/3 for first 12 edges
for i from 13 to 21 do mu[i,i+1]:=3: od:
#      and 3 for the last 9 edges
#
A_u:=Matrix(21,21,0):
for i from 1 to 21 do # diagonal entries of A_u
  A_u[i,i]:=-1/B*(mu[i,i]*i+mu[i,i+1]*(B-i)): od:
for i from 1 to 21 do # off-diagonal entries of A_u
  for k from 1 to i-1 do A_u[i,k]:=A_u[i,i]+mu[i,i]: od:
  for k from i+1 to 21 do A_u[i,k]:=A_u[i,i]+mu[i,i+1]:
od: od:
#
```

```
#
# We compute the eigenvalues:
with(LinearAlgebra):
evals:=evalf(Eigenvalues(A_u)):
evalm(evals);
```

$$
\begin{aligned}
&0.03033448065 + 0.3508691241i,\\
&-1.099781956 + 2.908783840i,\\
&-0.2847639055 + 0.2914327420i,\\
&-1.796376325 + 1.354320217i,\\
&-0.3937698928 + 0.2093609582i,\\
&-1.894085480 + 0.7212796463i,
\end{aligned}
$$

$$
\begin{aligned}
&-0.4453938101 + 0.1408808854i,\\
&-0.4715951453 + 0.08138816700i,\\
&-1.923518325 + 0.3232948910i,\\
&-0.4828957275 + 0.02663815111i,\\
&-1.930853280,\\
&-0.4828957275 - 0.02663815111i,\\
&-1.923518325 - 0.3232948910i,\\
&-0.4715951453 - 0.08138816700i,\\
&-0.4453938101 - 0.1408808854i,\\
&-1.894085480 - 0.7212796463i,\\
&-0.3937698928 - 0.2093609582i,\\
&-1.796376325 - 1.354320217i,\\
&-0.2847639055 - 0.2914327420i,\\
&-1.099781956 - 2.908783840i,\\
&0.03033448065 - 0.3508691241i
\end{aligned}
$$

```
#
# Largest real part of an eigenvalue is positive:
max(Re(evals));
```

$$0.03033448065$$

3. Computations for Lemma 2.39

In this section, we include the Maple code (with output) for demonstrating that call centre models with (at most) 4 customer classes are locally stable in critical load, when $q > 0$ (Lemma 2.39).

```
# We show here the that the four-customer-type
# critically loaded systems are locally stable.
# There are two essentially different arrangements
# of the four customer types:
# a   b   c   d
#  \ / \ / \ /
#   1   2   3
# and
# b--1
#     \
#      a--2--c
#     /
# d--3
# We do not divide through by B = sum_j beta_j
# in computations.
# This does not affect the sign of anything,
# and makes the results more legible.
#
```

```
# Case 1:
# a   b   c   d
#  \ / \ / \ /
#   1   2   3
# We compute the entries of A_c by first computing
# the entries of A_u (or, rather, of B*A_u as we
# explained above):
# Entries of A_u:
# Diagonal entries
#
Au1[aa]:=-mu[a1]*(beta[1]+beta[2]+beta[3]):
Au1[bb]:=-mu[b1]*beta[1]+mu[b2]*(beta[2]+beta[3]):
Au1[cc]:=-mu[c2]*(beta[1]+beta[2])+mu[c3]*beta[3]:
Au1[dd]:=-mu[d3]*(beta[1]+beta[2]+beta[3]):
#
# Off-diagonal entries
B:=beta[1]+beta[2]+beta[3]:
Au1[ab]:=0: Au1[ac]:=0: Au1[ad]:=0:
Au1[ba]:=Au1[bb]+B*mu[b1]:
  Au1[bc]:=Au1[bb]+B*mu[b2]:
  Au1[bd]:=Au1[bb]+B*mu[b2]:
Au1[ca]:=Au1[cc]+B*mu[c2]:
  Au1[cb]:=Au1[cc]+B*mu[c2]:
  Au1[cd]:=Au1[cc]+B*mu[c3]:
Au1[da]:=0: Au1[db]:=0: Au1[dc]:=0:
#
# Entries of A_c
```

```
cola:=1/4*(Au1[aa]+Au1[ba]+Au1[ca]+Au1[da]):
colb:=1/4*(Au1[ab]+Au1[bb]+Au1[cb]+Au1[db]):
colc:=1/4*(Au1[ac]+Au1[bc]+Au1[cc]+Au1[dc]):
cold:=1/4*(Au1[ad]+Au1[bd]+Au1[cd]+Au1[dd]):
Ac1[aa]:=Au1[aa]-cola: Ac1[ab]:=Au1[ab]-colb:
  Ac1[ac]:=Au1[ac]-colc: Ac1[ad]:=Au1[ad]-cold:
Ac1[ba]:=Au1[ba]-cola: Ac1[bb]:=Au1[bb]-colb:
  Ac1[bc]:=Au1[bc]-colc: Ac1[bd]:=Au1[bd]-cold:
Ac1[ca]:=Au1[ca]-cola: Ac1[cb]:=Au1[cb]-colb:
  Ac1[cc]:=Au1[cc]-colc: Ac1[cd]:=Au1[cd]-cold:
Ac1[da]:=Au1[da]-cola: Ac1[db]:=Au1[db]-colb:
  Ac1[dc]:=Au1[dc]-colc: Ac1[dd]:=Au1[dd]-cold:
#
# Matrix A_c
A[c,1]:=Matrix([[Ac1[aa], Ac1[ab], Ac1[ac], Ac1[ad]],
  [Ac1[ba], Ac1[bb], Ac1[bc], Ac1[bd]],
  [Ac1[ca], Ac1[cb], Ac1[cc], Ac1[cd]],
  [Ac1[da], Ac1[db], Ac1[dc], Ac1[dd]]]):
#
```

```
# Characteristic polynomial
# We know 0 is an eigenvalue, so we divide by x
with(LinearAlgebra):
charpol1:=simplify(CharacteristicPolynomial(A[c,1],x)/x):
#
```

```
# Coefficients and the expression c_2 c_1 - c_0.
# For stability, we need -c_2 > 0, c_1 > 0, -c_0 > 0,
# and -expr > 0.
# (The coefficients are called c[i,1], and the expression
# is called expr1, corresponding to case 1.)
#
c[2,1]:=-coeff(charpol1, x\symbol{94}2):
c[1,1]:=coeff(charpol1, x):
c[0,1]:=-coeff(charpol1, x, 0):
expr1:=expand(c[2,1]*c[1,1]-c[0,1]):
#
```

```
# c[i,1] and expr1 are polynomials in mu's and beta's.
# We will compute the sign of the minimal coefficient
# of -c_2, c_1, -c_0, and -expr as polynomials
# of mu's and beta's.
# Observing that these are all positive (+1), we see
# that the expressions are positive whenever mu's and
# beta's are positive (which they are).
# Thus, the matrix is stable.
#
signs1:=[sign(min(coeffs(-c[2,1]))),
  sign(min(coeffs(c[1,1]))),sign(min(coeffs(-c[0,1]))),
  sign(min(coeffs(-expr1)))];
#
#
```

$$signs1 := [1, 1, 1, 1]$$

```
# Case 2:
# b--1
#     \
#      a--2--c
#     /
# d--3
# We compute the entries of A_c by first computing
# the entries of A_u (or, rather, of B*A_u as we
# explained above):
# Entries of A_u:
# Diagonal entries
#
B:=beta[1]+beta[2]+beta[3]:
Au2[aa]:=-mu[a1]*beta[1]-mu[a2]*beta[2]-mu[a3]*beta[3]:
Au2[bb]:=-mu[b1]*B:
Au2[cc]:=-mu[c2]*B:
Au2[dd]:=-mu[d3]*B:
#
# Off-diagonal entries
Au2[ab]:=Au2[aa]+B*mu[a1]:
  Au2[ac]:=Au2[aa]+B*mu[a2]:
  Au2[ad]:=Au2[aa]+B*mu[a3]:
Au2[ba]:=0: Au2[bc]:=0: Au2[bd]:=0:
Au2[ca]:=0: Au2[cb]:=0: Au2[cd]:=0:
Au2[da]:=0: Au2[db]:=0: Au2[dc]:=0:
#
# Entries of A_c
cola:=1/4*(Au2[aa]+Au2[ba]+Au2[ca]+Au2[da]):
colb:=1/4*(Au2[ab]+Au2[bb]+Au2[cb]+Au2[db]):
colc:=1/4*(Au2[ac]+Au2[bc]+Au2[cc]+Au2[dc]):
cold:=1/4*(Au2[ad]+Au2[bd]+Au2[cd]+Au2[dd]):
Ac2[aa]:=Au2[aa]-cola: Ac2[ab]:=Au2[ab]-colb:
  Ac2[ac]:=Au2[ac]-colc: Ac2[ad]:=Au2[ad]-cold:
Ac2[ba]:=Au2[ba]-cola: Ac2[bb]:=Au2[bb]-colb:
  Ac2[bc]:=Au2[bc]-colc: Ac2[bd]:=Au2[bd]-cold:
Ac2[ca]:=Au2[ca]-cola: Ac2[cb]:=Au2[cb]-colb:
  Ac2[cc]:=Au2[cc]-colc: Ac2[cd]:=Au2[cd]-cold:
Ac2[da]:=Au2[da]-cola: Ac2[db]:=Au2[db]-colb:
  Ac2[dc]:=Au2[dc]-colc: Ac2[dd]:=Au2[dd]-cold:
#
# Matrix A_c
A[c,2]:=Matrix([[Ac2[aa], Ac2[ab], Ac2[ac], Ac2[ad]],
  [Ac2[ba], Ac2[bb], Ac2[bc], Ac2[bd]],
  [Ac2[ca], Ac2[cb], Ac2[cc], Ac2[cd]],
  [Ac2[da], Ac2[db], Ac2[dc], Ac2[dd]]]):
#
```

```
# Characteristic polynomial
# We know 0 is an eigenvalue, so we divide by x
charpol2:=simplify(CharacteristicPolynomial(A[c,2],x)/x):
#
```

```
# Coefficients and the expression c_2 c_1 - c_0.
# For stability, we need -c_2 > 0, c_1 > 0, -c_0 > 0,
# and -expr > 0.
# (The coefficients are called c[i,2], and the expression
# is called expr2, corresponding to case 2.)
#
c[2,2]:=-coeff(charpol2, x^2):
c[1,2]:=coeff(charpol2, x):
c[0,2]:=-coeff(charpol2, x, 0):
expr2:=expand(c[2,2]*c[1,2]-c[0,2]):
#
```

```
# c[i,2] and expr2 are polynomials in mu's and beta's.
# We will compute the sign of the minimal coefficient
# of -c_2, c_1, -c_0, and -expr as polynomials
# of mu's and beta's.
# Observing that these are all positive (+1), we see
# that the expressions are positive whenever mu's and
# beta's are positive (which they are).
# Thus, the matrix is stable.
#
signs2:=[sign(min(coeffs(-c[2,2]))),
  sign(min(coeffs(c[1,2]))),sign(min(coeffs(-c[0,2]))),
  sign(min(coeffs(-expr2)))];
#
```

$$signs2 := [1, 1, 1, 1]$$

4. Vertices of the level set of the Lyapunov function in §3.8

Here we describe the polyhedron $P \equiv \{X : \mathcal{L}(X) = 1\}$ constructed in §3.8 in the course of proving stability of the limit order book with 5 equally sized bins restricted to the interval $[\frac{1}{5} + \epsilon, \frac{4}{5} - \epsilon]$. Identifying $(x, y, z) \equiv (X(2), X(3), X(4))$, the (filled-in) polyhedron $\tilde{P} \equiv \{X : \mathcal{L}(X) \leq 1\}$ consists of the points satisfying the inequalities

$$|x| + |y| + |z| \leq 1, \qquad x - \frac{4}{5}y - \frac{9}{5}z \leq 1$$
$$\frac{4}{3}x + y + \frac{2}{3}z \leq 1, \qquad -2x - 3y - 4x \leq 1$$

and one of the four orthant constraints

$$x \geq 0, y \geq 0, z \geq 0 \ (+++) \qquad x \geq 0, y \geq 0, z \leq 0 \ (++-)$$
$$x \geq 0, y \leq 0, z \leq 0 \ (+--) \qquad x \leq 0, y \leq 0, z \leq 0 \ (---)$$

The polyhedron P has fifteen vertices

$\{0,0,0\}$, $\{0,1,0\}$, $\{0,0,1\}$, $\{\frac{1}{2},0,\frac{1}{2}\}$, $\{\frac{45}{58},\frac{2}{29},-\frac{9}{58}\}$, $\{\frac{6}{7},-\frac{1}{7},0\}$,
$\{\frac{29}{34},-\frac{2}{17},-\frac{1}{34}\}$, $\{\frac{3}{4},0,0\}$, $\{\frac{11}{50},\frac{6}{25},-\frac{27}{50}\}$, $\{0,\frac{3}{7},-\frac{4}{7}\}$, $\{\frac{11}{26},-\frac{6}{13},-\frac{3}{26}\}$,
$\{\frac{2}{5},-\frac{3}{5},0\}$, $\{0,-\frac{1}{3},0\}$, $\{-\frac{1}{2},0,0\}$, $\{0,0,-\frac{1}{4}\}$

and ten faces (defined as ordered sets of vertices, possibly with varying orientations; some of the faces will be non-convex)

$\{4,3,2\}$, $\{5,2,10,9\}$, $\{7,6,12,11\}$, $\{1,3,4,8\}$, $\{1,8,6,12,14\}$,
$\{1,3,2,10,15\}$, $\{1,15,14\}$, $\{7,5,9,11\}$, $\{2,4,8,6,7,5\}$,
$\{9,10,15,14,12,11\}$

The last three faces are the red, green, and blue face in Figure 3.1, which we reproduce below.