STOCHASTIC MODELS FOR SERVICE SYSTEMS AND LIMIT ORDER BOOKS

A Thesis
Presented to
The Academic Faculty

by

Xuefeng Gao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2013

STOCHASTIC MODELS FOR SERVICE SYSTEMS AND LIMIT ORDER BOOKS

Approved by:

Professor Jim Dai, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Ton Dieker, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Shijie Deng
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Anton J. Kleywegt
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr Mark S. Squillante
Stochastic Processes and Applications Group
IBM T.J. Watson Research Center

Date Approved: August 5, 2013

ACKNOWLEDGEMENTS

I am deeply grateful to all people who have helped and inspired me during my doctoral study in Georgia Tech.

First of all, I would like to thank my advisors, Dr Jim Dai and Dr Ton Dieker. Their passion, knowledge, and rigorous attitude toward research have deeply influenced me. They have been great mentors, providing me with tremendous support, encouragement and motivation. I am honored to have both of them as my advisors, and I feel very lucky to have the opportunity to work with them on exciting problems.

I would like to thank Dr Shijie Deng, Dr Anton J. Kleywegt, and Dr Mark S. Squillante for serving on my thesis committee. Dr Shijie Deng and Dr Anton J. Kleywegt have provided lots of ideas and inspiration on the third part of my thesis. Dr Mark S. Squillante introduced to me interesting business research problems and served as my mentor during my internships at IBM Thomas J. Watson Research Center.

I would like to thank Dr Gary Parker, Dr Paul Kvam and Ms Pam Morrison from the ISyE graduate studies office. I appreciate all the help you give me.

I would like to thank Gu Yu from Columbia University, who introduced to me hydrodynamic limit of interacting particle systems. It turns out to be a very useful tool in the third of my thesis.

I would like to thank all my friends. Your friendship have given me a great support during my studies. I enjoy having fun with you.

Finally, I would like to thank my parents and my girlfriend Jingjing, for their love and being with me all along this journey.

TABLE OF CONTENTS

PART III LIMIT ORDER BOOKS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Stochastic fluctuations can have profound impacts on engineered systems. Nonetheless, we can achieve significant benefits such as cost reduction based upon expanding our fundamental knowledge of stochastic systems. The primary goal of this thesis is to contribute to our understanding by developing and analyzing stochastic models for specific types of engineered systems. The knowledge gained can help management to optimize decision making under uncertainty.

This thesis has three parts. In Part I, we study many-server queues that model large-scale service systems such as call centers. We focus on the positive recurrence of piecewise Ornstein-Uhlenbeck (OU) processes and the validity of using these processes to predict the steady-state performance of the corresponding many-server queues. In Part II, we investigate diffusion processes constrained to the positive orthant under infinitesimal changes in the drift. This sensitivity analysis on the drift helps us understand how changes in service capacities at individual stations in a stochastic network would affect the steady-state queue-length distributions. In Part III, we study the trading mechanism known as limit order book. We are motivated by a desire to better understand the interplay between order book shape and and optimal executions. The goal is to characterize the temporal evolution of order book shape on the "macroscopic" time scale.

The contributions of Part I consist primarily of three aspects. First, we prove that the piecewise Ornstein-Uhlenbeck (OU) process arising from diffusion approximations of many-server queues is positive recurrent and has a unique stationary distribution. Second, we determine simple conditions for the existence of common quadratic Lyapunov functions, which are the key technical ingredients for Part I. Third, we establish

an interchange of limit theorem for many-server queues with customer abandonment. Our work is the first to rigorously justify that the stationary distribution of piecewise OU process is a good approximation of the steady-state behavior of the original queue with phase-type service requirements and exponential patience time distributions. The insights we obtain could be potentially used to determine the staffing level in service systems such as call centers to achieve certain service level objectives.

The contributions of Part II are twofold. First, we prove that any constrained function obtained from a Skorohod reflection map with oblique reflection, together with its (left) drift-derivative, is the unique solution to an augmented Skorohod problem. Second, we use this characterization to establish a basic adjoint relationship for the stationary distribution of the constrained diffusion process jointly with its left-derivative process. This work has the potential to lead to new numerical methods in the context of optimization and sensitivity analysis for queueing networks.

The contributions of Part III are also two-fold. First, we use the expected order flow parameters to give a "macroscopic" description of the order book shape dynamics, whose order book event-level description is a multi-dimensional continuous-time Markov chain. Second, we perform experiments to test our theoretical model against order book data from NYSE Arca. The initial empirical results suggest that our model could potentially predict the order book shape evolution reasonably well for highly liquid stocks in a relatively stationary environment. The knowledge gained could be potentially helpful for traders to design algorithms to optimally execute orders and thus reduce transaction costs.

STOCHASTIC MODELS FOR SERVICE SYSTEMS AND LIMIT ORDER BOOKS

PART I

Many-server queues

by

Xuefeng Gao

CHAPTER I

INTRODUCTION

1.1 Overview

Large-scale service systems, such as call centers, are becoming increasingly complex. These systems typically have a large amount of daily traffic with significant stochastic variability. In contrast with communication networks and manufacturing systems, customers in service systems could leave the system without service if their waiting time exceeds their patience time. This phenomenon of customer abandonment may significantly affect system performance. See Aksin et al. [3], Gans et al. [43] for surveys on call center management and related research questions.

For managers of service systems, one important decision is to determine how many servers should be used to achieve service level objectives, such as answering 70% of calls within 2 minutes. To make these staffing decisions, one has to understand, model and analyze those systems. Many-server queueing models turn out to be useful to gain insights into the operation and design of large-scale service systems with hundreds or even thousands of servers (agents).

Despite past and foreseeable advances in computer hardware and architectures, except for the simplest cases, the sheer size of such many-server queueing systems prohibits exact analysis. In particular, when the service times are random with an arbitrary distribution, there is still no general analytical or numerical tool to efficiently and accurately predict the steady-state performance of such many-server queueing systems. Computer simulation is often the only remaining tool available, but it can be slow when the system is heavily-loaded and the number of servers is large.

Pioneered by Halfin and Whitt [50], diffusion approximations have been used for

performance analysis of heavily-loaded many-server queueing systems. In diffusion approximations of many-server queues, one typically scales the space of a sequence of queueing systems and sends the traffic intensity of the systems to one at a suitable rate. The main appeal of diffusion approximations is its simplicity: the diffusion model (process) can be specified using a drift coefficient and a diffusion coefficient. Thus it provides a relatively tractable and rigorous approximation for the (stochastic) queue length and waiting time dynamics in service systems.

Recently, Dai and He [25] proposed diffusion models to approximate a $GI/Ph/n+M$ system. In a $GI/Ph/n+M$ queue, there are n servers. The interarrival times of customers are independently and identically distributed (i.i.d.) following a general distribution (GI). The service times are i.i.d. following a phase-type distribution (Ph). The set of phase-type distributions is dense in the field of all positive-valued distributions. Hence it can be used to approximate any general service time distribution. In addition, customers patience times are i.i.d. following an exponential distribution ($+M$). The approximations in Dai and He [25] are rooted in many-server heavy traffic limit theorems proved in Dai et al. [24]. The numerical examples in Dai and He [25] demonstrate that the steady-state performance of the diffusion model provides a remarkably accurate estimate for the steady-state performance of the corresponding queueing system, even when the number of servers is moderate (e.g., 20 servers).

Our goal in the first part of this thesis is to provide a solid mathematical foundation for the diffusion approximation procedure in Dai and He [25]. We address two specific questions: (1) positive recurrence of a class of multi-dimensional diffusion processes known as piecewise Ornstein-Uhlenbeck (OU) processes; and (2) validity of the heavy-traffic steady-state approximations for $GI/Ph/n+M$ queues. Piecewise OU processes have piecewise linear drift coefficient and they arise from diffusion approximations of many-server queues with phase-type service requirements, see Puhalskii

and Reiman [87] and Dai et al. [24]. We show in Chapter 3 that the piecewise OU process is positive recurrent and has a unique stationary distribution under some natural conditions. In addition, we prove in Chapter 4 an interchange of limit theorem which rigorously justifies that the stationary distribution of the piecewise OU process is a good approximation of the stationary distribution of the original many-server queue.

We now discuss the challenges in proving the positive recurrence of piecewise OU processes. A standard technique for proving stability of queueing systems is to first establish the stability of a so-called fluid model and then to appeal to general theory for establishing stochastic stability (see, e.g., Dupuis and Williams [36], Dai [22], Stolyar [100]). However, this theory is restricted to systems with nonnegative fluid levels which are attracted to the origin. The fluid analog of a piecewise Ornstein-Uhlenbeck process does not possess this property.

As an alternative to the fluid model framework, the family of *quadratic* Lyapunov functions is a natural choice for establishing positive recurrence. Piecewise OU processes exhibit different behavior in two regions of the state space, corresponding to 'overload' and 'underload'. The two regions are separated by a hyperplane, which corresponds to 'critical load'. In each of the two regions, a piecewise OU process can be thought of as a first-order linear differential equation with stochastic noise. A standard technique in proving its positive recurrence is to use a quadratic Lyapunov function to prove stability of such first-order linear differential equations. However, the two different regions of a piecewise OU process pose considerable challenges to apply this methodology. A natural approach would be to 'paste together' two quadratic Lyapunov functions from the two regions, but our attempts in this direction have failed. In fact, it is well-known that a diffusion with two stable regimes can lead to an instable hybrid system, see Yin and Zhu [110] for related examples.

We next discuss the challenges in establishing the validity of heavy traffic steady-state approximations for $GI/Ph/n + M$ queues. When the number of servers n

is fixed, a $GI/Ph/n+M$ system can be represented by a certain continuous time Markov process $\Xi^n = \{\Xi^n(t) : t \geq 0\}$. Often $\Xi^n(t)$ converges in distribution to $\Xi^n(\infty)$ as time $t \to \infty$, where the random variable $\Xi^n(\infty)$ has the stationary distribution of Ξ^n. On the other hand, Dai et al. [24] show that $\tilde{\Xi}^n$, as a sequence of stochastic processes that are centered and scaled versions of Ξ^n, converges in distribution to some diffusion process $\tilde{\Xi}$ as $n \to \infty$ under a heavy traffic condition. The limit process $\tilde{\Xi}$ is a piecewise Ornstein-Uhlenbeck (OU) process. The convergence proved in [24] implies that each finite $t \geq 0$, $\tilde{\Xi}^n(t)$ converges in distribution to $\tilde{\Xi}(t)$, but, as in almost all diffusion limits, does not cover the case when $t = \infty$. Therefore, our goal is to show $\tilde{\Xi}^n(\infty)$ converges in distribution to $\tilde{\Xi}(\infty)$ as $n \to \infty$.

1.2 Contributions and related work

Our contributions in the first part of this thesis are threefold. First, we establish the positive recurrence of piecewise OU processes. Using the interpretation of the diffusion parameters in terms of a many-server queueing system, we prove the following results in Chapter 3: (1) For a slightly underloaded system without abandonment, we show that there exists a quadratic Lyapunov function which yields the desired positive recurrence using the Foster-Lyapunov criterion (Theorem 3.2). In general, this quadratic Lyapunov function is not explicit and non-unique. (2) We show that no quadratic Lyapunov function can satisfy the Foster-Lyapunov criterion for systems with abandonment. (3) We construct a suitable non-quadratic Lyapunov function to prove positive recurrence for systems with abandonment (Theorem 3.3).

The main building blocks for these two types of Lyapunov functions are so-called common quadratic Lyapunov functions (CQLFs), which are widely used in the theory of control. Such functions play an important role in the stability analysis for deterministic linear systems, with different dynamics in different parts of the state space (or, more generally, operating under a switching rule). They are called *common* quadratic

Lyapunov functions since they serve as a quadratic Lyapunov function in each part of the state space. There is a vast body of literature on CQLFs and related theory, see the survey Shorten et al. [96] for details. Although quadratic Lyapunov functions are ubiquitous in the literature on queueing systems (Dai and Prabhakar [27], Gamarnik and Momčilović [40], Tassiulas and Ephremides [101], and Dieker and Shin [33]), to our knowledge, our work here is the first to exploit CQLFs in this context.

Second, we determine simple conditions for the existence of common quadratic Lyapunov functions, which is of considerable interest as mentioned in the section on open problems of Shorten et al. [96]. Theorem 3.1 establishes such a result in the context of M-matrices and rank-1 perturbations. The theorem shows that existence of a CQLF is guaranteed after merely verifying that certain vectors are nonnegative. It is a first result of this kind. Its proof relies on a delicate analysis involving Chebyshev polynomials, as well as on an extension of the recent work of King and Nathanson [63] and Shorten et al. [95] summarized in Proposition 3.3.

Third, we prove an interchange limit theorem for $GI/Ph/n+M$ queues, where the arrival process is a renewal process, service time distribution is of phase-type and customer patience time distribution is exponential. We prove that $\tilde{\Xi}^n(\infty)$, which has the stationary distribution of the scaled state process $\tilde{\Xi}^n$ of $GI/Ph/n+M$ queues, converges in distribution to $\tilde{\Xi}(\infty)$ as $n \to \infty$; see Theorem 4.1 and Corollary 4.1 in Chapter 4 below. Often, one can also prove that the many-server diffusion limit $\tilde{\Xi}(t)$ converges in distribution to $\tilde{\Xi}(\infty)$ as $t \to \infty$. Formally, our results can be stated as

$$\lim_{n\to\infty} \lim_{t\to\infty} \tilde{\Xi}^n(t) \stackrel{d}{=} \lim_{t\to\infty} \lim_{n\to\infty} \tilde{\Xi}^n(t),$$

which is known as the interchange limit theorem.

There has been a surge of interest in establishing interchange limit theorems in the last ten years in both the conventional heavy traffic setting and many-server heavy traffic setting. To prove an interchange limit theorem, when the stationary distribution of the limit process $\tilde{\Xi}$ is unique, it is sufficient to prove that the sequence of

random variables $\{\tilde{\Xi}^n(\infty) : n \geq 1\}$ is tight. Gamarnik and Zeevi [42] pioneered an approach in proving the tightness in the context of generalized Jackson networks in conventional heavy traffic when all distributions are assumed to have finite exponential moments. Key to their proof is the construction of a geometric Lyapunov function. Inspired by this work, Budhiraja and Lee [16] devised an alternative method to prove the tightness along with some other results also in the context of generalized Jackson networks, but with the minimal two moment assumption on all distributions. Budhiraja and Lee [16] did not construct a geometric Lyapunov function, but they cleverly utilized and sharpened a fluid limit approach introduced in Dai and Meyn [26], and their approach is potentially more general and obtains sharper results.

We follow the approach in Gamarnik and Zeevi [42] by constructing a geometric Lyapunov function; see Lemma 4.3 in Section 4.4 below. We heavily rely on a delicate analysis of the behavior of our Lyapunov functions applied to a fluid model for $GI/Ph/n+M$ queues. In particular, when applied to the fluid model, we show that our Lyapunov function decreases at a rate that is proportional to the size of the fluid state when it is far away from origin; see part (b) of Lemma A.7. Because of the customer abandonment in our model, when the waiting fluid is high, the decreasing rate in our Lyapunov function should be expected due to abandonment. However, when the waiting fluid is not high, but a large fluid state is due to the huge imbalance of servers among different service phases, the decreasing rate is by no means obvious. Our proof relies critically on common quadratic Lyapunov functions. It remains an open problem whether the approach in Budhiraja and Lee [16] can be adapted to our setting.

In the many-server setting without customer abandonment, Halfin and Whitt [50] is the first paper to establish a many-server diffusion limit for the $GI/M/n$ model. In the same paper they proved the tightness result. Gamarnik and Momčilović [40] prove the tightness result where service time distribution is lattice-valued with a

finite support. Gamarnik and Goldberg [39] have generalized this result to $GI/GI/n$ queues. In a single class, multiple server pool model, Tezcan [102] proved asymptotic optimality of some routing policies and some interchange limit results.

In the many-server setting with customer abandonment, Gamarnik and Stolyar [41] proved a tightness result. In their model, customers have many classes. Each class has its own homogeneous Poisson arrival process, exponential service time distribution, and exponential patience time distribution with class dependent rate. The service policy in choosing which class of customers to serve next can be arbitrary as long as it is non-idling. When the service policy is first-in-first-out across customer classes and the patience rate is independent of customer classes, their model reduces to a special case of our model considered in this thesis. Their proof critically relies on the assumption that service time distributions are exponential and it remains an open problem whether their tightness result holds for non-exponential service time distributions. Empirical study in Brown et al. [15] suggests that the service time distributions are approximately lognormal, not exponential.

In the conventional heavy traffic setting, Gurvich [49] systematically generalized the results in Gamarnik and Zeevi [42] to multiclass queueing networks. Katsuda [60] proves some interchange limit results for a multiclass single-server queue with feedback under various disciplines. Ye and Yao [109] studied interchange limit results in a head-of-line bandwidth sharing model with two customer classes and two servers.

1.3 Organization

The rest of Part I of this thesis is organized as follows. Chapter 2 reviews many-server queues that allow customer abandonment and their diffusion approximations. Chapter 3 is devoted to the positive recurrence of piecewise OU processes. Chapter 4 focuses on the validity of interchange of heavy-traffic steady-state limit.

1.4 Notation

This section contains the notations used in Part I of this thesis.

All random variables and stochastic processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ unless otherwise specified. For a positive integer d, $\mathbb{R}^d$ denotes the d-dimensional Euclidean space. Given a subset S of some Euclidean space, the space of right-continuous functions $f : \mathbb{R}_+ \to S$ with left limits in $(0, \infty)$ is denoted by $\mathbb{D}(\mathbb{R}_+, S)$ or simply $\mathbb{D}(S)$. Each stochastic process with sample paths in $\mathbb{D}(S)$ is considered to be a $\mathbb{D}(S)$-valued random element. The space $\mathbb{D}(S)$ is assumed to be endowed with the Skorohod J_1-topology. Given $y \in \mathbb{D}(S)$ and $t > 0$, we set $\|y\|_t = \sup_{0 \le s \le t} |y(s)|$, where $|\cdot|$ denotes the Euclidean norm in S. For a sequence of random elements $\{X_n : n = 1, 2, \ldots\}$ taking values in a metric space, we write $X_n \Rightarrow X$ to denote the convergence of X_n to X in distribution. The space of functions $f : \mathbb{R}^K \to \mathbb{R}$ that are twice continuously differentiable is denoted by $C^2(\mathbb{R}^K)$. We use ∇ to denote the gradient operator. Given $x \in \mathbb{R}$, we set $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. Given a $K \times K$ matrix M, we use M' to denote its transpose, and similarly for vector transposition. We write M_{ij} for its (i, j)-th entry. We write $M > 0$ ($M < 0$) if M is a positive (negative) definite matrix and $M \ge 0$ ($M \le 0$) if it is a positive (negative) semi-definite matrix. Let the matrix norm of M be $|M| = \sum_{ij} |M_{i,j}|$, where $|M_{ij}|$ is the absolute value of M_{ij}. All vectors are envisioned as column vectors. For a K-dimensional vector u, we write $|u|$ for its Euclidean norm. For two K-dimensional vectors u and v, we write $u' \ge v'$ ($u' > v'$) if $u_k \ge v_k$ ($u_k > v_k$) for each $k = 1, 2, \ldots, K$. The inner product of u and v is denoted by $u'v$, which is $\sum_{k=1}^K u_k v_k$. We reserve I for the $K \times K$ identity matrix and e for the K-dimensional vector of ones.

CHAPTER II

$G/Ph/n+GI$ QUEUES AND DIFFUSION APPROXIMATIONS

In this chapter, we review many-server queues that allow for customer abandonment, and provide necessary background for validity of heavy traffic steady state approximation. The service time distribution in these queues is restricted to be phase-type. Phase-type distributions can be used to approximate any positive-valued distribution, see Neuts [78]. We first introduce the many-server queueing model and then describe diffusion approximations for performance analysis of these systems in the so-called QED regime. Most of the materials of this chapter can be found in Dai et al. [24].

2.1 $G/Ph/n+GI$ queues and piecewise OU processes

In this section we present background on $G/Ph/n+GI$ queues and piecewise OU (Ornstein-Uhlenbeck) processes. Their connections will be made clear in Subsection 2.2.3. In a $G/Ph/n+GI$ queue, there are n identical servers. The arrival process $E=\{E(t), t\geq 0\}$ is assumed to be general (the first G), where $E(t)$ denotes the number of customer arrivals to the system by time t. Upon his arrival to the system, a customer gets into service immediately if there is an idle server; otherwise, he waits in a waiting buffer that holds a first-in-first-out (FIFO) queue. The buffer size is assumed to be infinite. When a server finishes a service, the server removes the leading customer from the waiting buffer and starts to serve the next customer; when the queue is empty, the server begins to idle.

In this model, each customer has a patience time: when a customer's waiting time in queue exceeds his patience time, the customer abandons the system without any service. We assume that the patience times of customers who arrive after time 0,

form a sequence of i.i.d. random variables that have a general distribution (the last $+GI$). We assume the distribution function F of patience time satisfies

$$F(0) = 0, \quad \text{and} \quad \alpha = \lim_{x \to 0+} x^{-1} F(x) < \infty. \tag{2.1.1}$$

The service times are *i.i.d.* random variables, following a phase-type distribution. Examples of phase-type distributions include exponential distributions, Erlang distributions and Hyper-exponential distributions. The precise definition is introduced in the next subsection.

2.1.1 Phase-type distributions

Let p be a K-dimensional nonnegative vector whose entries sum up to one, ν be a K-dimensional positive vector, and P be a $K \times K$ sub-stochastic matrix. We assume that the diagonal entries of P are zero and $\mathsf{I} - P$ is invertible. Consider a continuous-time Markov chain with $K+1$ phases (or states) where phases $1, 2 \ldots, K$ are transient and phase $K+1$ is absorbing. The initial distribution of the Markov chain is p. The amount of time it stays in phase k is exponentially distributed with mean $\frac{1}{\nu_k}$. When it leaves phase k, the Markov chain enters phase j with probability P_k^j or enters phase $K+1$ with probability $1 - \sum_{j \le K} P_k^j$. The rate matrix of this transient markov chain is

$$\hat{G} = \begin{pmatrix} \hat{F} & \hat{c} \\ 0 & 0 \end{pmatrix}$$

where $\hat{F} = \mathrm{diag}(\nu)(P - \mathsf{I})$ is a $K \times K$ matrix and $\hat{c} = -\hat{F}e$.

Definition 2.1 *A continuous phase-type random variable with parameters (p, ν, P) is defined to be the first time until the continuous-time Markov chain with initial distribution p and rate matrix $\hat{G}$ reaches state $K+1$.*

This completes the introduction of the many-server queueing model: a $G/Ph/n+GI$ queue. In particular, when the arrival process is a renewal process and the patience

times are *i.i.d.* following an exponential distribution, the resulting system is denoted by $GI/Ph/n+M$ (the first GI signifies renewal arrivals and the last $+M$ signifies exponential patience time). We study an interchange of limit theorem for this queue in Chapter 4.

2.1.2 Piecewise OU processes

In this subsection, we introduce piecewise OU (Ornstein-Uhlenbeck) processes. They are special diffusion processes. We first provide some background on diffusion processes.

Let $\{W(t)\}$ be a standard Brownian motion in any dimension. A K-dimensional diffusion process Y is the strong solution to a stochastic differential equation of the form

$$dY(t) = b(Y(t))dt + \sigma(Y(t))dW(t),$$

where the drift coefficient $b(\cdot)$ and the diffusion coefficient $\sigma(\cdot)$ have appropriate sizes and satisfy the following Lipschitz continuity condition: there exists some $C > 0$ such that

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq C|x - y| \quad \text{for all } x, y \in \mathbb{R}^K. \tag{2.1.2}$$

For a real-valued function $V \in C^2(\mathbb{R}^K)$, the generator G of Y applied to V is given by, for $y \in \mathbb{R}^K$,

$$GV(y) = (\nabla V(y))' b(y) + \frac{1}{2} \sum_{i,j} (\sigma\sigma')_i^j(y) \frac{\partial^2 V}{\partial y_i \partial y_j}(y). \tag{2.1.3}$$

We refer to Rogers and Williams [90, Chapter V] for more details on diffusion processes.

A key concept in defining piecewise OU processes is M-matrices, which we introduce below. We call a matrix nonnegative when each element of the matrix is nonnegative.

Definition 2.2 (M-matrix) *A $K \times K$ matrix $\hat{R}$ is said to be an M-matrix if it can be expressed as $\hat{R} = s\mathsf{I} - \mathsf{N}$ for some $s > 0$ and some nonnegative matrix N with the property that $r(\mathsf{N}) \leq s$, where $r(\mathsf{N})$ is the spectral radius of N. The matrix $\hat{R}$ is nonsingular if $r(\mathsf{N}) < s$.*

We are now ready to define piecewise Ornstein-Uhlenbeck (OU) processes.

Definition 2.3 (Piecewise OU processes) *Let $\hat{p}$ be a K-dimensional probability vector, e be the K-dimensional vector of ones, and let $\hat{R}$ be a $K \times K$ nonsingular M-matrix. For $\hat{\alpha}, \hat{\beta} \in \mathbb{R}$, a K-dimensional diffusion process Y is called a piecewise Ornstein-Uhlenbeck (OU) process if it has drift coefficient*

$$b(y) = -\hat{\beta}\hat{p} - \hat{R}(y - \hat{p}(e'y)^+) - \hat{\alpha}\hat{p}(e'y)^+, \tag{2.1.4}$$

and diffusion coefficient $\sigma(y) \equiv \sigma$ for all $y \in \mathbb{R}^K$, such that $\sigma\sigma'$ is a $K \times K$ nonsingular matrix.

As in Dai et al. [24], we call this process a *piecewise* OU process since the drift coefficient is affine (hence OU process) yet it differs on each side of the hyperplane $\{y \in \mathbb{R}^K : e'y = 0\}$ (hence piecewise). Indeed, for $e'y \geq 0$ we have $b(y) = -\hat{\beta}\hat{p} - \hat{R}(\mathsf{I} - \hat{p}e')y - \hat{\alpha}\hat{p}(e'y)$ while for $e'y \leq 0$ we have $b(y) = -\hat{\beta}\hat{p} - \hat{R}y$. In conjunction with $\sigma(y) \equiv \sigma$, this implies the Lipschitz continuity condition (2.1.2). As a consequence, the piecewise OU process Y is well-defined as a diffusion process.

The quantities $\hat{\alpha}, \hat{\beta}, \hat{R}, \hat{p}$ on the right-hand side of (2.1.4) have natural interpretation in the context of $G/Ph/n + GI$ queues. This will be illustrated in the next subsection. We also mention recent work Pang and Yao [80] and Gurvich [48], where piecewise OU processes find useful applications in many-server queueing network with switchovers and continuous time Markovian queues.

Throughout Part I of the this thesis, we impose the following assumption.

Assumption 2.1 *Each component of the row vector $e'\hat{R}$ is nonnegative, i.e.,*

$$e'\hat{R} \geq 0'.$$

2.2 Diffusion approximations

In this section, we introduce the diffusion approximations of $G/Ph/n+GI$ queues for performance analysis in the so-called QED regime. We first introduce the QED regime.

2.2.1 QED regime

Consider a sequence of $G/Ph/n+GI$ queues indexed by n. For the nth system, write E^n for the arrival process and λ^n for the arrival rate. The arrival rate $\lambda^n \to \infty$ when $n \to \infty$, while the service time and the patience time distributions do not change with n. Define

$$\rho^n = \frac{\lambda^n}{n\mu} \quad \text{and} \quad \beta^n = \sqrt{n}(1-\rho^n).$$

The quantity ρ^n is said to be the traffic intensity of the nth system. We assume that

$$\lim_{n\to\infty} \beta^n = \beta \quad \text{for some } \beta \in \mathbb{R}. \tag{2.2.1}$$

When condition (2.2.1) is satisfied, the sequence of systems is critically loaded, and is said to be in the Quality-and-Efficiency-Driven (QED) regime or the Halfin-Whitt regime.

To facilitate analysis below, we write

$$\hat{E}^n(t) = E^n(t) - \lambda^n t$$

and we assume

$$\frac{1}{\sqrt{n}}\hat{E}^n(\cdot) \Rightarrow \tilde{E}(\cdot), \tag{2.2.2}$$

where $\tilde{E}$ is a Brownian motion, and the symbol " $\Rightarrow$ " means weak convergence.

2.2.2 State processes

To describe the "state" of the system as time evolves, we let $Z_k^n(t)$ denote the number of customers *in phase k service* in the nth system at time t; service times in phase k are exponentially distributed with rate ν_k. We use $Z^n(t)$ to denote the corresponding

K-dimensional vector. We call $Z^n = \{Z^n(t), t \geq 0\}$ the *server-allocation process*. Let $N^n(t)$ denote the number of customers in the nth system at time t, either in queue or in service. Setting

$$X^n(t) = N^n(t) - n \quad \text{for } t \geq 0, \tag{2.2.3}$$

we call $X^n = \{X^n(t), t \geq 0\}$ the *total-customer-count process* in the nth system. One can check that $(X^n(t))^+$ is the number of customers waiting in queue at time t, and $(X^n(t))^-$ is the number of idle servers at time t. Clearly,

$$e'Z^n(t) = n - (X^n(t))^- \quad \text{for } t \geq 0.$$

In addition, suppose that each customer, including those initial customers who are waiting in the buffer at time zero, samples his first service phase that he is yet to enter following distribution p at his arrival time. One can stratify the customers in the buffer according to their first service phases. For $j = 1, 2, \cdots, K$, we use $Q_j^n(t)$ to denote the number of waiting customers at time t whose initial service phase is phase j. If phase j is not a first service phase for any customer, we take $Q_j^n(t) = 0$. We use $Y_j^n(t)$ to denote the number of phase j customers in the system at time t, i.e.,

$$Y_j^n(t) = Q_j^n(t) + Z_j^n(t).$$

Let $Q^n(t)$ and $Y^n(t)$ denote the corresponding d-dimensional vectors. For $t \geq 0$, set

$$\tilde{Y}^n(t) = \frac{1}{\sqrt{n}}(Y^n(t) - n\gamma), \tag{2.2.4}$$

$$\tilde{Z}^n(t) = \frac{1}{\sqrt{n}}(Z^n(t) - n\gamma), \tag{2.2.5}$$

$$\tilde{X}^n(t) = \frac{1}{\sqrt{n}}X^n(t), \tag{2.2.6}$$

where γ is given by

$$\gamma = \mu R^{-1} p, \tag{2.2.7}$$

$$R = (\mathsf{I} - P')\,\mathrm{diag}(\nu). \tag{2.2.8}$$

One can check that $\sum_{k=1}^{K} \gamma_k = 1$, and one interprets γ_k to be the fraction of phase k load on the n servers.

2.2.3 Convergence to the diffusion limit

In this section, we describe the many-server diffusion limits for $G/Ph/n+GI$ queues, where the limiting diffusion process is a piecewise OU process. We first introduce some notations. Let c_a^2 be the squared coefficient of variation of interarrival time distribution. Set matrices $H[0]$ and $H[j]$ as follows:

$$H[0]_k^l = \begin{cases} p_k(1-p_l) & \text{if } k = l \\ -p_k p_l & \text{otherwise} \end{cases} \quad \text{and} \quad H[j]_k^l = \begin{cases} P_j^k(1-P_j^l) & \text{if } k = l \\ -P_j^k P_j^l & \text{otherwise} \end{cases} \tag{2.2.9}$$

for $j = 1, 2, \ldots, K$.

In addition, recall that α is the abandonment rate defined in (2.1.1), β is the slack in the arrival rate relative to a critically loaded system as defined in (2.2.1), while p and R are the parameters of the phase-type service-time distribution. Now we are ready to state the diffusion limits for many-server queues that allow abandonment, see [24].

Theorem 2.1 *Consider a sequence of $G/Ph/n+GI$ queues satisfying (2.2.1) and (2.2.2). Assume that we have $(\tilde{X}^n(0), \tilde{Z}^n(0)) \Rightarrow (\tilde{X}(0), \tilde{Z}(0))$ for a pair of random variables $(\tilde{X}(0), \tilde{Z}(0))$. Then*

$$(\tilde{X}^n, \tilde{Y}^n, \tilde{Z}^n) \Rightarrow (\tilde{X}, \tilde{Y}, \tilde{Z}) \quad \text{as } n \to \infty, \tag{2.2.10}$$

where the process $\tilde{Y}$ is a K-dimensional piecewise OU process with drift coefficient

$$b(y) = -\beta p - R(y - p(e'y)^+) - \alpha(e'y)^+, \tag{2.2.11}$$

and nonsingular diffusion coefficient σ with

$$\Sigma = \sigma\sigma' = \mu(c_a^2 pp' + H[0]) + \sum_{j=1}^{d} \nu_j \gamma_j H[j] + (\mathsf{I} - P')diag(\nu)diag(\gamma)(\mathsf{I} - P), \tag{2.2.12}$$

where matrix $H[0]$ and $H[j]$ are given in (2.2.9). Moreover, $(\tilde{X}, \tilde{Z})$ is a (K+1)-dimensional degenerate continuous time Markov processes satisfying

$$\tilde{X}(t) = e'\tilde{Y}(t) \quad \textit{and} \quad \tilde{Z}(t) = \tilde{Y}(t) - p\tilde{X}(t)^{+} \quad \textit{for all } t \geq 0. \tag{2.2.13}$$

We now discuss Assumption 2.1 in the context of $G/Ph/n + GI$ queues. We compare (2.2.11) and (2.1.4) since the many-server diffusion limit $\tilde{Y}$ in (2.2.10) is a particular piecewise OU process. The matrix R defined in (2.2.8) takes the form of $(\mathsf{I} - P')\,\mathrm{diag}\{\nu\}$, where P is a transient matrix. Transience of P corresponds to customers who eventually leave the system after receiving a sufficient amount of service. This implies that $e'R = e'(\mathsf{I} - P')\,\mathrm{diag}\{\nu\} \geq 0$. Therefore, we conclude that in this setting R is a nonsingular M-matrix and that Assumption 2.1 is satisfied.

CHAPTER III

STABILITY OF PIECEWISE OU PROCESSES

In this chapter, we prove that the general piecewise OU process introduced in Subsection 2.1.2 is positive recurrent and has a unique stationary distribution under some natural conditions. Since the many-server diffusion limit $\tilde{Y}$ in Theorem 2.1 is a particular piecewise OU process, $\tilde{Y}$ has a unique stationary distribution. This stability property is important since it provides a solid mathematical foundation for numerical algorithms to compute stationary distribution of piecewise OU process in Dai and He [25].

The key technical ingredients for proving stability of piecewise OU process are common quadratic Lyapunov functions (CQLFs), which are widely used in the theory of control. Such functions play an important role in the stability analysis for deterministic linear systems, with different dynamics in different parts of the state space (or, more generally, operating under a switching rule). In Theorem 3.1, we determine simple conditions for the existence of CQLFs in the context of M-matrices and rank-1 perturbations. It is a first result of this kind.

Using the interpretation of the diffusion parameters in terms of a many-server queueing system, our stability results in this chapter can be summarized as follows: (1) For a slightly underloaded system without abandonment, we show that there exists a quadratic Lyapunov function which yields the desired positive recurrence using the Foster-Lyapunov criterion (Theorem 3.2). In general, this quadratic Lyapunov function is not explicit and non-unique. (2) We construct a suitable non-quadratic Lyapunov function to prove positive recurrence for systems with abandonment (Theorem 3.3).

The rest of the chapter is organized as follows. We first discuss some background on positive recurrence and Lyapunov functions. Section 3.2 is devoted to common quadratic Lyapunov functions, which are the key technical tools. The main results on stability are presented in Section 3.3. Section 3.4 is dedicated to the proofs.

3.1 Positive recurrence and Lyapunov functions

In this section, we recall the definitions and the criteria for positive recurrence and exponential ergodicity in the context of general diffusion processes.

Let $\mathbb{E}_\pi$ be the expectation operator with respect to a probability distribution π.

Definition 3.1 (Positive recurrence and stationary distribution) *A K-dimensional diffusion process Y is positive recurrent if for any $y \in \mathbb{R}^K$ and any compact set C in $\mathbb{R}^K$ with positive Lebesgue measure, we have*

$$\mathbb{E}(\tau_C | Y(0) = y) < \infty,$$

where $\tau_C = \inf\{t \geq 0 : Y(t) \in C\}$ is the hitting time of the set C. We call a probability distribution π on $\mathbb{R}^K$ a stationary distribution for Y if for every bounded continuous function f: $\mathbb{R}^K \to \mathbb{R}$,

$$\mathbb{E}_\pi[f(Y(t))] = \mathbb{E}_\pi[f(Y(0))] \quad \textit{for all } t \geq 0.$$

In the following, we assume that the diffusion coefficient of the diffusion process Y is uniformly nonsingular. That is, there exists some $c \in (0, \infty)$ such that for all $y \in \mathbb{R}^K$ and $a \in \mathbb{R}^K$,

$$a'\sigma(y)\sigma(y)'a \geq ca'a. \tag{3.1.1}$$

The next result gives a sufficient criterion for positive recurrence of diffusion processes, see Khasminskii [62, Section 3.7, 4.3 and 4.4], and Meyn and Tweedie [77, Section 4]. Uniqueness of the stationary distribution follows from Peszat and Zabczyk [82] in view of condition (3.1.1).

Proposition 3.1 (Foster-Lyapunov criterion) *Let Y be a diffusion process satisfying (3.1.1). Suppose that there exists a nonnegative function $V \in C^2(\mathbb{R}^K)$ and some $r > 0$ such that, for any $|y| > r$,*

$$GV(y) \leq -1.$$

In addition, suppose that $V(y) \to \infty$ as $|y| \to \infty$. Then Y is positive recurrent and has a unique stationary distribution. The function V is called a Lyapunov function.

We now introduce the concept of exponential ergodicity. For any positive measurable function $f \geq 1$ and any signed measure m, we write $||m||_f = \sup_{|g| \leq f} |m(g)|$.

Definition 3.2 (Exponential ergodicity) *Suppose that the diffusion process Y is positive recurrent and that it has a unique stationary distribution π. Given a function $f \geq 1$, we say that Y is f-exponentially ergodic if there exists a $c \in (0, 1)$ and a real-valued function B such that for all $t > 0$ and $y \in \mathbb{R}^K$,*

$$||P^t(y, \cdot) - \pi(\cdot)||_f \leq B(y)c^t,$$

where P^t is the transition function of Y. If $f \equiv 1$, we simply say that Y is exponentially ergodic.

For $f \geq 1$, we have $||P^t(y, \cdot) - \pi(\cdot)||_1 \leq ||P^t(y, \cdot) - \pi(\cdot)||_f$, and we deduce that f-exponential ergodicity implies exponential ergodicity. The following result gives a criterion for exponential ergodicity, see Meyn and Tweedie [77, Section 6].

Proposition 3.2 *Suppose that Y is a diffusion process with a unique stationary distribution. If there is a nonnegative function $V \in C^2(\mathbb{R}^K)$ such that $V(y) \to \infty$ as $|y| \to \infty$ and for some $c > 0, d < \infty$,*

$$GV(y) \leq -cV(y) + d \quad \text{for any } y \in \mathbb{R}^K,$$

then Y is $(V + 1)$-exponentially ergodic.

3.2 Common quadratic Lyapunov functions

In this section we introduce common quadratic Lyapunov functions (CQLFs). Such functions play a central role in the stability analysis of deterministic switched linear systems, which is discussed in Section 3.2.2. We use CQLFs as building blocks to construct Lyapunov functions to prove positive recurrence of piecewise OU processes. At this point it is best to distinguish CQLFs for switched linear systems from Lyapunov functions in the context of the Foster-Lyapunov criterion.

3.2.1 Background and definitions

Quadratic Lyapunov functions form a cornerstone of stability theory for ordinary differential equations. Consider the linear system $\dot{y}(t) = By(t)$ where $y(t) \in \mathbb{R}^K$, $B \in \mathbb{R}^{K\times K}$ is a fixed real matrix and $\dot{y}(t)$ is the derivative of y with respect to t. For $Q \in \mathbb{R}^{K\times K}$, the quadratic form L given by $L(y) = y'Qy$ for $y \in \mathbb{R}^K$ is called a quadratic Lyapunov function for the matrix B if Q is positive definite and $QB + B'Q$ is negative definite. In this case, there exists a constant $C > 0$ such that

$$\frac{d}{dt}L(y(t)) = y(t)'(QB + B'Q)y(t) \leq -CL(y(t)) < 0 \quad \text{for all } t \geq 0,$$

and thus we can conclude that $L(y(t)) \leq e^{-Ct}L(y(0))$. This implies that $L(y(t)) \to 0$ as $t \to \infty$, thus $y(t) \to 0$ as $t \to \infty$. It is a standard fact in Lyapunov stability theory that the existence of a quadratic Lyapunov function L is equivalent to all eigenvalues of B having negative real part, see, e.g., Berman and Plemmons [9, Section 6.2].

The following definition, tailored to our setting in order to allow for a singular matrix, plays an important role in our analysis. Other versions can be found in Shorten and Narendra [97] and Shorten et al. [96]. Recall that an eigenvalue of a matrix is called (geometrically) simple if its corresponding eigenspace is one-dimensional.

Definition 3.3 (CQLF) *Let $B_1 \in \mathbb{R}^{K\times K}$ have all eigenvalues with negative real part and let $B_2 \in \mathbb{R}^{K\times K}$ have all eigenvalues with negative real part except for a simple zero*

eigenvalue. For $Q \in \mathbb{R}^{K\times K}$, the quadratic form L given by $L(y) = y'Qy$ for $y \in \mathbb{R}^K$ is called a common quadratic Lyapunov function (CQLF) for the pair (B_1, B_2) if Q is positive definite and

$$QB_1 + B_1'Q < 0,$$

$$QB_2 + B_2'Q \leq 0.$$

3.2.2 The CQLF existence problem

The CQLF existence problem for a pair of matrices has its roots in the study of stability criteria for switched linear systems. These systems have the form $\dot{y}(t) = B(\tau)y(t)$ where $B(\tau) \in \{B_1, B_2\}$ with $B_i \in \mathbb{R}^{K\times K}$ for $i = 1, 2$ and where the switching function τ may depend on both y and t. The existence of a CQLF for the pair (B_1, B_2) guarantees that all solutions of the systems are bounded under arbitrary switching function τ. The CQLF existence problem is also closely related to the Kalman-Yacubovich-Popov lemma in the development of adaptive control algorithms and the Lur'e problem in nonlinear feedback analysis. For more details consult Kalman [57], Boyd et al. [14] and the recent survey paper by Shorten et al. [96]. For an arbitrary matrix pair, no simple analytic and verifiable conditions are known for the pair to admit a CQLF. In the special case where the difference of the matrices has rank one, King and Nathanson [63] shows that if both B_1 and B_2 are Hurwitz, i.e., all eigenvalues of the matrices B_1, B_2 have negative real part, then there exists a positive definite matrix Q such that $QB_1 + B_1'Q < 0$ and $QB_2 + B_2'Q < 0$ if and only if the matrix product B_1B_2 has no real negative eigenvalues. Note that in this case, both B_1 and B_2 are nonsingular. A similar CQLF existence result has been obtained in Shorten et al. [95] when one of the matrices (B_1 or B_2) is singular.

We now state a result on the CQLF existence problem for a pair of matrices with one of them being singular. It is essentially the main theorem in [95] but we relax their assumptions. Let $B \in \mathbb{R}^{K\times K}$ be a real matrix and let $g, h \in \mathbb{R}^K$. The

proposition below is stated in Shorten et al. [95] under the assumptions that (B, g) is controllable, meaning that the vectors $g, Bg, B^2g, \ldots$ span $\mathbb{R}^K$, and that (B, h) is observable, meaning that the vectors $h, B'h, (B')^2h, \ldots$ span $\mathbb{R}^K$. Using techniques from King and Nathanson [63], we show that these assumptions are unnecessary and we state the result in its full generality here. A proof is given in Appendix A.1.

Proposition 3.3 *Suppose that all eigenvalues of matrix B have negative real part and all eigenvalues of $B - gh'$ have negative real part, except for a simple zero eigenvalue. Then there exists a CQLF for the pair $(B, B - gh')$ if and only if the matrix product $B(B - gh')$ has no real negative eigenvalues and a simple zero eigenvalue.*

3.3 Stability results

In this section, we present our results on positive recurrence of the general piecewise OU process Y in Definition 2.3 in Section 2.1.2. Note the many-server diffusion limit $\tilde{Y}$ in (2.2.10) is a particular piecewise OU process. Key to our stability results is the following theorem, which uses Proposition 3.3 to establish the existence of a CQLF for certain matrix pairs. Recall from Definition 2.3 that $\hat{R}$ is a nonsingular M-matrix, $\hat{p}$ is a probability vector, and e is a vector of ones. Note that we are working under Assumption 2.1.

Theorem 3.1 *There exists a CQLF for both the pair $(-\hat{R}, -\hat{R}(\mathsf{I} - \hat{p}e'))$ and the pair $(-\hat{R}, -(\mathsf{I} - \hat{p}e')\hat{R})$.*

By Theorem 3.1, there exists a CQLF L for the pair $(-\hat{R}, -\hat{R}(\mathsf{I} - \hat{p}e'))$ and another CQLF $\widetilde{L}$ for the pair $(-\hat{R}, -(\mathsf{I} - \hat{p}e')\hat{R})$. Typically there are many CQLFs corresponding to these pairs, i.e., L and $\widetilde{L}$ are not unique. Note that L and $\widetilde{L}$ are closely related in the following sense. If the CQLF L for the pair $(-\hat{R}, -\hat{R}(\mathsf{I} - pe'))$ is given by $L(y) = y'Qy$ for some $Q > 0$ and for all $y \in \hat{\mathbb{R}}^K$, then one readily checks that the quadratic form $\widetilde{L}$ given by $\widetilde{L}(y) = y'(\hat{R}'Q\hat{R})y$ for $y \in \mathbb{R}^K$ is a CQLF for the

pair $(-\hat{R}, -(\mathsf{I} - \hat{p}e')\hat{R})$. We remark that, apart from special cases, the CQLFs from Theorem 3.1 are not explicit.

We know from Theorem 3.1 that there exists a CQLF L for the pair $(-\hat{R}, -\hat{R}(\mathsf{I} - \hat{p}e'))$, where L is given by $L(y) = y'Qy$ for some $Q > 0$ and for all $y \in \mathbb{R}^K$. We are able to use the quadratic form L as a Lyapunov function in the Foster-Lyapunov criterion of Proposition 3.1 to prove the following result.

Theorem 3.2 *If $\hat{\alpha} = 0$ and $\hat{\beta} > 0$, then the piecewise OU process Y in Definition 2.3 is positive recurrent and has a unique stationary distribution.*

For $\hat{\alpha} > 0$, no quadratic function can serve as a Lyapunov function in the Foster-Lyapunov criterion to prove positive recurrence of the piecewise OU process Y, see Appendix A.2 for details. Despite this fact, still relying on Theorem 3.1, we overcome this difficulty by constructing a suitable non-quadratic Lyapunov function. Specifically, there exists a CQLF $\widetilde{L}$ for the pair $(-\hat{R}, -(\mathsf{I} - \hat{p}e')\hat{R})$ by Theorem 3.1, where $\widetilde{L}$ is given by $\widetilde{L}(y) = y'\widetilde{Q}y$ for some $\widetilde{Q} > 0$ and for all $y \in \mathbb{R}^K$. A suitable approximation to the function f, given by for all $y \in \mathbb{R}^K$,

$$f(y) = (e'y)^2 + \kappa\widetilde{L}(y - \hat{p}(e'y)^+) \quad \text{for some large constant } \kappa,$$

provides the desired non-quadratic Lyapunov function in the Foster-Lyapunov criterion to prove positive recurrence of Y when $\hat{\alpha} > 0$. Note that, in queueing terminology, the vector $y - \hat{p}(e'y)^+$ relates to the customers in service, and not to those in the buffer. We therefore need the extra term $(e'y)^2$. Applying Proposition 3.2 with the same non-quadratic Lyapunov function yields exponential ergodicity of Y for $\hat{\alpha} > 0$. We use a smooth approximation of f as a Lyapunov function in the Foster-Lyapunov criterion of Proposition 3.1 instead of using f directly since $f \notin C^2(\mathbb{R}^K)$. This leads to the following result.

Theorem 3.3 *If $\hat{\alpha} > 0$, then the piecewise OU process Y in Definition 2.3 is positive recurrent and has a unique stationary distribution. Moreover, Y is exponentially ergodic.*

3.4 Proof of stability results

3.4.1 Proof of Theorem 3.1

Proof of Theorem 3.1 We only establish the existence of a CQLF for the pair $(-\hat{R}, -\hat{R}(\mathsf{I}-\hat{p}e'))$, since the existence of a CQLF for the other pair $(-\hat{R}, -(\mathsf{I}-\hat{p}e')\hat{R})$ follows directly. Since $-\hat{R} - (-\hat{R}(\mathsf{I} - \hat{p}e')) = -\hat{R}\hat{p}e'$ is a rank-one matrix, in view of Proposition 3.3, we need to check three conditions:

(a) All eigenvalues of $-\hat{R}$ have negative real part;

(b) All eigenvalues of $-\hat{R}(\mathsf{I} - \hat{p}e')$ have negative real part except for a simple zero eigenvalue;

(c) The matrix product $\hat{R}^2(\mathsf{I} - \hat{p}e')$ has no real negative eigenvalues and a simple zero eigenvalue.

We first prove (a) and (b). It is known that all eigenvalues of a nonsingular M-matrix have positive real part, and all eigenvalues of a singular M-matrix have nonnegative real part, see Berman and Plemmons [9, Chapter 6]. Since $\hat{R}$ is a nonsingular M-matrix, we immediately get (a). For (b), it is clear that $-\hat{R}(\mathsf{I} - \hat{p}e')$ has a simple zero eigenvalue. We notice that $(\mathsf{I} - \hat{p}e')\hat{R} = \hat{R} - \hat{p}e'\hat{R}$ where $e'R \geq 0'$ by Assumption 2.1, $\hat{p}$ is a nonnegative vector and $\hat{R}$ is a nonsingular M-matrix, so the off-diagonal elements of $(\mathsf{I} - \hat{p}e')\hat{R}$ are nonpositive. Using this in conjunction with the fact that both $\mathsf{I} - \hat{p}e'$ and $\hat{R}$ are M-matrices, we find that $(\mathsf{I} - \hat{p}e')\hat{R}$ is also an M-matrix and all its eigenvalues have nonnegative real part, see Berman and Plemmons [9, Exercise 5.2]. Thus we get (b) after a similarity transform.

We now concentrate on proving (c). The key ingredient of the proof is an identity for Chebyshev polynomials. Suppose that $\hat{R}^2(\mathsf{I} - \hat{p}e')$ has a real negative eigenvalue $-\lambda$ with $\lambda > 0$, and write v for the corresponding left eigenvector, thus we have $v'\hat{R}^2(\mathsf{I} - \hat{p}e') = -\lambda v'$. Right-multiplying by $\hat{p}$ on both sides, we obtain $v'\hat{p} = 0$ and the following equality:

$$0 = v'\hat{R}^2(\mathsf{I} - \hat{p}e') + \lambda v' = v'\hat{R}^2(\mathsf{I} - \hat{p}e') + \lambda v'(\mathsf{I} - \hat{p}e') = v'(\hat{R}^2 + \lambda\mathsf{I})(\mathsf{I} - \hat{p}e'). \quad (3.4.1)$$

Since $\hat{R}$ is a nonsingular M-matrix having only eigenvalues with positive real part, the matrix $(\hat{R}^2 + \lambda\mathsf{I})$ is invertible for all $\lambda > 0$. Also, by the fact that $\hat{p}$ is a nonnegative probability vector with $e'\hat{p} = 1$, we deduce the matrix $(\mathsf{I} - \hat{p}e')$ has an eigenvalue 0 and the corresponding left eigenvector must be in the form of ce' for some $c \neq 0$. Thus, it follows from (3.4.1) that $v' = ce'(\hat{R}^2 + \lambda\mathsf{I})^{-1}$ for some $c \neq 0$. We show below that $e'(\hat{R}^2 + \lambda\mathsf{I})^{-1}$ is a positive vector for all $\lambda > 0$, i.e,

$$e'(\hat{R}^2 + \lambda\mathsf{I})^{-1} > 0' \quad \text{for all } \lambda > 0. \quad (3.4.2)$$

This yields a contradiction in view of $v'\hat{p} = 0$. By definition of a nonsingular M-matrix, $\hat{R}$ is of the form $s\mathsf{I} - \mathsf{N}$, where N is a nonnegative matrix with spectral radius smaller than s and $e'\hat{R} \geq 0$ by Assumption 2.1. Equation (3.4.2) thus states that for all $\lambda > 0$ and for every nonnegative matrix N with spectral radius smaller than s and $se' \geq e'\mathsf{N}$,

$$e'((s\mathsf{I} - \mathsf{N})^2 + \lambda\mathsf{I})^{-1} > 0'.$$

Equivalently, we show the following inequality: for all $y \in (0, 1)$ and for every nonnegative matrix N with $r(\mathsf{N}) < 1$ and $e' \geq e'\mathsf{N}$,

$$e'(y(\mathsf{I} - \mathsf{N})^2 + (1 - y)\mathsf{I})^{-1} > 0'. \quad (3.4.3)$$

Therefore, to show (c), it suffices to prove (3.4.3) for fixed N and $y \in (0, 1)$.

Our strategy to prove (3.4.3) is to use a matrix series expansion and connections with Chebyshev polynomials. Chebyshev polynomials of the second kind U_n can be

defined by the following trigonometric form:

$$U_n(\cos\vartheta) = \frac{\sin(n+1)\vartheta}{\sin\vartheta} \quad \text{for } n = 0, 1, 2, 3, \ldots. \tag{3.4.4}$$

Moreover, for $z \in [-1, 1]$ and $t \in (-1, 1)$, the generating function of U_n is

$$\sum_{n=0}^{\infty} U_n(z)t^n = \frac{1}{1 - 2tz + t^2}. \tag{3.4.5}$$

Refer to [2, Chapter 22] for more details. The scalar version of the left-hand side of (3.4.3) admits the following expansion: for $x, y \in (0, 1)$,

$$\frac{1}{y(1-x)^2 + 1 - y} = \sum_{n=0}^{\infty} C_n(y)x^n, \tag{3.4.6}$$

where $C_n(y) = U_n(\sqrt{y})(\sqrt{y})^n$ for all $n \geq 0$. This can readily be verified with (3.4.5). In particular, we have

$$C_0(y) = U_0(y) \equiv 1 \quad \text{for all } y \in (0, 1). \tag{3.4.7}$$

For fixed $y \in (0, 1)$, the radius of convergence of the power series in (3.4.6) is larger than 1. Since $r(\mathsf{N}) < 1$, we immediately obtain that for $y \in (0, 1)$,

$$(y(\mathsf{I} - \mathsf{N})^2 + (1 - y)\mathsf{I})^{-1} = \sum_{n=0}^{\infty} C_n(y)\mathsf{N}^n. \tag{3.4.8}$$

Let $y \in (0, 1)$ be fixed and define ϑ through $\sqrt{y} = \cos\vartheta \in (0, 1)$. Using the trigonometric form (3.4.4) of U_n, we can then show by induction that for any $m \geq 1$,

$$\begin{aligned} \sum_{n=1}^{m} C_n(y) &= \sum_{n=1}^{m} U_n(\sqrt{y})(\sqrt{y})^n \\ &= \sum_{n=1}^{m} \frac{\sin(n+1)\vartheta}{\sin\vartheta} \cdot (\cos\vartheta)^n \\ &= \frac{\cos^2\vartheta}{\sin^2\vartheta}[1 - (\cos\vartheta)^{m-1} \cdot \cos(m+1)\vartheta] > 0. \end{aligned} \tag{3.4.9}$$

Since N is nonnegative and $e' \geq e'\mathsf{N}$, we immediately get $e'\mathsf{N}^n \geq e'\mathsf{N}^{n+1} \geq 0$ for all $n \geq 0$. Combining this fact with (3.4.9), we obtain

$$e' \sum_{n=1}^{k} C_n(y)\mathsf{N}^n \geq \sum_{n=1}^{k} C_n(y)e'\mathsf{N}^k \geq 0' \quad \text{for all } k \geq 1. \tag{3.4.10}$$

Therefore, from (3.4.7), (3.4.8) and (3.4.10) we conclude that for all $y \in (0,1)$,

$$\begin{aligned} e'((1-y)\mathsf{I} + y(\mathsf{I}-\mathsf{N})^2)^{-1} &= e' \sum_{n=0}^{\infty} C_n(y)\mathsf{N}^n \\ &= \lim_{k\to\infty} e' \sum_{n=1}^{k} C_n(y)\mathsf{N}^n + e' \\ &\geq 0' + e' = e' > 0'. \end{aligned}$$

This concludes the proof of (c) and we deduce that there exists a CQLF for the pair $(-\hat{R}, -\hat{R}(\mathsf{I} - \hat{p}e'))$.

To prove the existence of a CQLF for the other pair $(-\hat{R}, -(\mathsf{I}-\hat{p}e')\hat{R})$, we note that $-(\mathsf{I}-pe')\hat{R}$ has the same spectrum as $-\hat{R}(\mathsf{I}-\hat{p}e')$ and the matrix product $\hat{R}(\mathsf{I}-pe')\hat{R}$ has the same spectrum as $\hat{R}^2(\mathsf{I} - \hat{p}e')$. Application of Proposition 3.3 completes the proof of Theorem 3.1.

3.4.2 Proof of Theorem 3.2

In this section we prove Theorem 3.2. Key to the proof is the CQLF constructed from Theorem 3.1.

Proof of Theorem 3.2: If $\hat{\alpha} = 0$, then from (2.1.4) we know that Y has the piecewise linear drift

$$b(y) = -\hat{\beta}\hat{p} - \hat{R}(y - \hat{p}(e'y)^+).$$

By Theorem 3.1, there exists a CQLF

$$L(y) = y'Qy, \tag{3.4.11}$$

where Q is a positive definite matrix such that

$$Q(-\hat{R}) + (-\hat{R})'Q < 0, \tag{3.4.12}$$

$$Q(-\hat{R}(\mathsf{I} - \hat{p}e')) + (-(\mathsf{I} - e\hat{p}')\hat{R}')Q \leq 0. \tag{3.4.13}$$

We claim that given any positive constant $C > 0$, there exists a constant $M > 0$ such that if $|y| > M$,

$$(\nabla L(y))' b(y) \leq -C. \tag{3.4.14}$$

We discuss the cases $e'y < 0$ and $e'y \geq 0$ separately.

Case 1: $e'y < 0$. In this case, we have

$$(\nabla L(y))' b(y) = y'[Q(-\hat{R}) + (-\hat{R})'Q]y - 2\hat{\beta}\hat{p}'Qy.$$

By (3.4.12), the quadratic term dominates if $|y|$ is large. Thus there exists a constant $M_1 > 0$ such that when $e'y < 0$ and $|y| > M_1$,

$$(\nabla L(y))' b(y) \leq -C.$$

Case 2: $e'y \geq 0$. In this case, we have

$$(\nabla L(y))' b(y) = y'[Q(-\hat{R}(\mathsf{I} - \hat{p}e')) + (-(\mathsf{I} - e\hat{p}')\hat{R}')Q]y - 2\hat{\beta}\hat{p}'Qy. \tag{3.4.15}$$

To overcome the difficulty caused by the singularity of $-\hat{R}(\mathsf{I} - \hat{p}e')$, we decompose y as follows:

$$y = a\hat{p} + \xi, \tag{3.4.16}$$

where $\xi'\hat{p} = 0$ and $a \in \mathbb{R}$. Then we have

$$|y|^2 = |a\hat{p}|^2 + |\xi|^2 \quad \text{and} \quad e'y = a + e'\xi \geq 0. \tag{3.4.17}$$

Note that $\hat{p}'[Q(-\hat{R}(\mathsf{I} - \hat{p}e')) + (-(\mathsf{I} - e\hat{p}')\hat{R}')Q]\hat{p} = 0$. Using (3.4.13), we obtain $\hat{p}'[Q(-\hat{R}(\mathsf{I}-\hat{p}e'))+(-(\mathsf{I}-e\hat{p}')\hat{R}')Q] = 0'$. This immediately implies $\hat{p}'[Q(-\hat{R}(\mathsf{I}-\hat{p}e')] = 0$. Since $(\mathsf{I} - \hat{p}e')$ has a simple zero eigenvalue, we have

$$\hat{p}'Q = be'\hat{R}^{-1} \quad \text{for some } b \neq 0.$$

Using this fact, we rewrite the left-hand side of (3.4.13) as follows:

$$\begin{aligned} &Q(-\hat{R}(\mathsf{I} - \hat{p}e')) + (-(\mathsf{I} - e\hat{p}')\hat{R}')Q \\ &\quad = ((\mathsf{I} - e\hat{p}')\hat{R}') \cdot (-Q\hat{R}^{-1} - (\hat{R}^{-1})'Q) \cdot (\hat{R}(\mathsf{I} - \hat{p}e')). \end{aligned} \tag{3.4.18}$$

After left-multiplying by $(\hat{R}^{-1})'$ and right-multiplying by $\hat{R}^{-1}$ in (3.4.12), we deduce that $[-Q\hat{R}^{-1} - (\hat{R}^{-1})'Q]$ is a negative definite matrix. Moreover, since $\xi'\hat{p} = 0$, from (3.4.16) and (3.4.18) we know that there exists some $c > 0$ such that

$$\begin{aligned} & y'[Q(-\hat{R}(\mathsf{I} - \hat{p}e')) + (-(\mathsf{I} - e\hat{p}')\hat{R}')Q]y \\ &= y'[(\mathsf{I} - e\hat{p}')\hat{R}' \cdot (-Q\hat{R}^{-1} - (\hat{R}^{-1})'Q) \cdot \hat{R}(\mathsf{I} - \hat{p}e')]y \\ &= \xi'((\mathsf{I} - e\hat{p}')\hat{R}') \cdot (-Q\hat{R}^{-1} - (\hat{R}^{-1})'Q) \cdot (\hat{R}(\mathsf{I} - \hat{p}e'))\xi \\ &\leq -c|\xi|^2. \end{aligned} \tag{3.4.19}$$

Therefore, from (3.4.15) we have that for any y with $e'y \geq 0$,

$$(\nabla L(y))'b(y) \leq -c|\xi|^2 - 2\hat{\beta}\hat{p}'Q\xi - 2\hat{\beta}a\hat{p}'Q\hat{p} \tag{3.4.20}$$

$$\leq -c|\xi|^2 - 2\hat{\beta}\hat{p}'Q\xi + 2\hat{\beta}\hat{p}'Q\hat{p}e'\xi, \tag{3.4.21}$$

where the second inequality is obtained from (3.4.17), $\hat{\beta} > 0$ and $\hat{p}'Q\hat{p} > 0$. For $|y|$ large, if $|\xi| \geq r$ for some large constant r, we obtain $(\nabla L(y))'b(y) \leq -C$ since the quadratic term $-c|\xi|^2$ in (3.4.21) dominates. If $|\xi| < r$ and $|y|$ large, we deduce from (3.4.17) that a must be positive and large, i.e.,

$$a \geq \frac{1}{|\hat{p}|}\sqrt{|y|^2 - r^2}.$$

Hence the dominating term in (3.4.20) is $-2\hat{\beta}a\hat{p}'Q\hat{p}$ and we immediately obtain $(\nabla L(y))'b(y) \leq -C$ whenever $|y|$ is large. Therefore, there exists a constant $M_2 > 0$ such that when $e'y \geq 0$ and $|y| > M_2$,

$$(\nabla L(y))'b(y) \leq -C.$$

On setting $M = \max\{M_1, M_2\}$, we immediately get (4.5.1).

Now set $C = |\sum_{i,j} Q_{ij}(\sigma\sigma')_{ij}| + 1$. Equations (3.4.11) and (4.5.1) imply that for $|y| > M$,

$$GL(y) = \sum_{i,j} Q_{ij}(\sigma\sigma')_{ij} + (\nabla L(y))'b(y) \leq -1.$$

The proof of Theorem 3.2 is complete after applying Proposition 3.1.

3.4.3 Proof of Theorem 3.3

In this section we prove Theorem 3.3. Throughout this section, C is a generic positive constant which may differ from line to line but is independent of y.

By Theorem 3.1, there exists a positive definite matrix $\widetilde{Q}$ with $|\widetilde{Q}| = 1$ such that

$$\widetilde{Q}(-\hat{R}) + (-\hat{R})'\widetilde{Q} < 0, \tag{3.4.22}$$

$$\widetilde{Q}(-(\mathsf{I} - \hat{p}e')\hat{R}) + (-\hat{R}'(\mathsf{I} - e\hat{p}'))\widetilde{Q} \leq 0. \tag{3.4.23}$$

We construct a non-quadratic Lyapunov function $V \in C^2(\mathbb{R}^K)$ as follows. Let

$$V(y) = (e'y)^2 + \kappa[y - \hat{p}\phi(e'y)]'\widetilde{Q}[y - \hat{p}\phi(e'y)], \tag{3.4.24}$$

where κ is a positive constant to be decided later and $\phi(x)$ is a real-valued $C^2(\mathbb{R})$ function, approximating $x \mapsto x^+$. Specifically, fix $\epsilon > 0$ and let

$$\phi(x) = \begin{cases} x & \text{if } x \geq 0, \\ -\frac{1}{2}\epsilon & \text{if } x \leq -\epsilon, \\ \text{smooth} & \text{if } -\epsilon < x < 0. \end{cases}$$

We piece $x \geq 0$ and $x \leq -\epsilon$ together in a smooth way such that ϕ is in $C^2(\mathbb{R})$, $-\frac{1}{2}\epsilon \leq \phi(x) \leq x^+$ and $0 \leq \dot{\phi}(x) \leq 1$ for any $x \in \mathbb{R}$, where $\dot{\phi}$ is the derivative of ϕ. This function ϕ evidently exists. Note that $V \in C^2(\mathbb{R}^K)$, but that it is not a CQLF due to its non-quadratic nature. We summarize the key result in the following proposition, which implies Theorem 3.3.

Proposition 3.4 *If $\hat{\alpha} > 0$, there exists a constant $C > 0$ such that when $|y|$ is large enough, we have*

$$(\nabla V(y))'b(y) \leq -C|y|^2 \quad \textit{and} \quad \left|\frac{\partial^2 V}{\partial y_i \partial y_j}(y)\right| \leq C|y| \quad \textit{for any } i, j. \tag{3.4.25}$$

Consequently, when $|y|$ is large,

$$GV(y) \leq -C|y|^2 \leq -1. \tag{3.4.26}$$

Proof We first study $(\nabla V(y))'b(y)$. From (3.4.24), we have for all $y \in \mathbb{R}^K$,

$$(\nabla V(y))' = 2(e'y)e' + 2\kappa(y' - \hat{p}'\phi(e'y))\widetilde{Q}[\mathsf{I} - \hat{p}e'\dot{\phi}(e'y)]. \tag{3.4.27}$$

We separately discuss the cases $e'y \geq 0$, $e'y \leq -\epsilon$ and $-\epsilon < e'y < 0$.

Case 1: $e'y \geq 0$. In this case, let $x = e'y$ and $z = y - \hat{p}x = (\mathsf{I} - \hat{p}e')y$, then we have

$$\begin{aligned}(\nabla V(y))'b(y) &= [2(e'y)e' + 2\kappa y'(\mathsf{I} - e\hat{p}')\tilde{Q}(\mathsf{I} - \hat{p}e')](-\hat{R}(\mathsf{I} - \hat{p}e')y - \hat{\alpha}\hat{p}e'y - \hat{\beta}\hat{p}) \\ &= -2\hat{\alpha}x^2 - \kappa z'[\widetilde{Q}(\mathsf{I} - \hat{p}e')\hat{R} + \hat{R}'(\mathsf{I} - e\hat{p}')\widetilde{Q}]z - 2x\hat{\beta} - 2xe'\hat{R}z.\end{aligned}$$

Suppose we have shown that there exists $C > 0$ such that

$$z'[\widetilde{Q}(\mathsf{I} - \hat{p}e')\hat{R} + \hat{R}'(\mathsf{I} - e\hat{p}')\widetilde{Q}]z \geq C|z|^2, \tag{3.4.28}$$

we then obtain that

$$(\nabla V(y))'b(y) \leq -2\hat{\alpha}x^2 - \kappa C|z|^2 - 2x\hat{\beta} - 2xe'\hat{R}z.$$

Since $\hat{\alpha} > 0$, we can select $\kappa > 0$ large so that $\frac{1}{2}(2\hat{\alpha}x^2 + \kappa C|z|^2) > 2|xe'\hat{R}z|$ for any (x, z), where κ is independent of (x, z) or y. Then we have,

$$(\nabla V(y))'b(y) \leq -\hat{\alpha}x^2 - \frac{1}{2}\kappa C|z|^2 - 2x\hat{\beta}.$$

Note that $|y| = |px + z| \leq C|(x, z)|$, so that $|(x, z)|$ is large whenever $|y|$ is large. We conclude that for $|y|$ large,

$$\begin{aligned}(\nabla V(y))'b(y) &\leq -C|(x, z)|^2 \\ &\leq -C|y|^2.\end{aligned}$$

It remains to prove (3.4.28). We use a similar argument as for (3.4.19). Observe that

$$(\hat{R}^{-1}\hat{p})'[\widetilde{Q}(\mathsf{I} - \hat{p}e')\hat{R} + \hat{R}'(\mathsf{I} - e\hat{p}')\widetilde{Q}](\hat{R}^{-1}\hat{p}) = 0,$$

which implies that $\widetilde{Q}\hat{R}^{-1}\hat{p} = be$ for some $b \in \mathbb{R}$. Thus, we obtain

$$\begin{aligned} & z'[\widetilde{Q}(\mathsf{I} - \hat{p}e')\hat{R} + \hat{R}'(\mathsf{I} - e\hat{p}')\widetilde{Q}]z \\ &= z'\hat{R}'(\mathsf{I} - e\hat{p}')[(\hat{R}^{-1})'\widetilde{Q} + \widetilde{Q}\hat{R}^{-1}](\mathsf{I} - \hat{p}e')\hat{R}z. \end{aligned} \tag{3.4.29}$$

Since $\hat{R}$ is a nonsingular M-matrix, $\hat{R}^{-1}$ is a nonnegative matrix (Berman and Plemmons [9, Chapter 6]) and we deduce that

$$e'\hat{R}^{-1}\hat{p} > 0. \tag{3.4.30}$$

This implies that $(\mathsf{I} - \hat{p}e')\hat{R}z \neq 0$ since $e'z = e'(\mathsf{I} - \hat{p}e')y = 0$ in this case. From (3.4.22) we know that $(\hat{R}^{-1})'\widetilde{Q} + \widetilde{Q}\hat{R}^{-1}$ is a positive definite matrix. Now (3.4.28) follows from (3.4.29).

Case 2: $e'y < -\epsilon$. In this case, we have $\phi(e'y) = -\frac{1}{2}\epsilon$ and $\dot{\phi}(e'y) = 0$. From (3.4.22), there exists $C > 0$ such that

$$\begin{aligned} (\nabla V(y))'b(y) &= (2(e'y)e' + 2\kappa y'\widetilde{Q} + \kappa\epsilon\hat{p}'\widetilde{Q})(-\hat{R}y - \hat{\beta}\hat{p}) \\ &= -2\kappa\left[y'(\widetilde{Q}\hat{R} + \hat{R}'\widetilde{Q})y + \frac{1}{2}(\epsilon\hat{p}'\widetilde{Q}\hat{R} + \hat{\beta}\hat{p}'\widetilde{Q})y + \frac{1}{2}\epsilon\hat{\beta}\hat{p}'\widetilde{Q}p\right] \\ &\quad -2e'y \cdot (e'\hat{R}y + \hat{\beta}) \\ &\leq -2\kappa\left[C|y|^2 + \frac{1}{2}(\epsilon\hat{p}'\widetilde{Q}\hat{R} + \hat{\beta}\hat{p}'\widetilde{Q})y + \frac{1}{2}\epsilon\hat{\beta}\hat{p}'\widetilde{Q}\hat{p}\right] - 2e'y \cdot (e'\hat{R}y + \hat{\beta}) \\ &\leq -2\kappa\left[C|y|^2 + \frac{1}{2}(\epsilon\hat{p}'\widetilde{Q}\hat{R} + \hat{\beta}p'\widetilde{Q})y + \frac{1}{2}\epsilon\hat{\beta}\hat{p}'\widetilde{Q}\hat{p}\right] + \kappa C(|y|^2 + |y|) \\ &\leq -\kappa(C|y|^2 - C|y| - C), \end{aligned}$$

where κ is again chosen to be independent of y, but large enough such that $|2e'y \cdot (e'\hat{R}y + \hat{\beta})| < \kappa C(|y|^2 + |y|)$. Thus for $|y|$ large and $e'y < -\epsilon$, we have

$$(\nabla V(y))'b(y) \leq -C|y|^2.$$

Case 3: $-\epsilon \leq e'y \leq 0$. In this case we use the property that $0 \leq \dot{\phi}(e'y) \leq 1$.

Note that we have

$$\begin{aligned}
&(\nabla V(y))'b(y)\\
&= (2(e'y)e' + 2\kappa(y' - \hat{p}'\phi(e'y))\widetilde{Q}(\mathsf{I} - \hat{p}e'\dot{\phi}(e'y)))(-\hat{R}y - \hat{\beta}\hat{p})\\
&= 2e'ye'(-\hat{R}y - \hat{\beta}\hat{p})\\
&\quad +2\kappa\dot{\phi}(e'y)(y' - \hat{p}'\phi(e'y))\widetilde{Q}(\mathsf{I} - \hat{p}e')(-\hat{R}y - \hat{\beta}\hat{p})\\
&\quad +2\kappa(1 - \dot{\phi}(e'y))(y' - \hat{p}'\phi(e'y))\widetilde{Q}(-\hat{R}y - \hat{\beta}\hat{p}).
\end{aligned}$$

We write

$$y = a\hat{R}^{-1}\hat{p} + \xi,$$

where ξ is orthogonal to $\hat{R}^{-1}\hat{p}$ and $a \in \mathbb{R}$, so that

$$|y|^2 = ca^2 + |\xi|^2, \quad \text{for some } c > 0. \tag{3.4.31}$$

From (3.4.30), we have $e'\hat{R}^{-1}\hat{p} > 0$. Without loss of generality we assume that $e'\hat{R}^{-1}\hat{p} = 1$. Then $e'y = a + e'\xi$ and we get

$$\begin{aligned}
&(\nabla V(y))'b(y)\\
&= -2(a + e'\xi)(\hat{\beta} + e'\hat{R}\xi + a)\\
&\quad +\kappa\dot{\phi}(e'y)(\xi'[\widetilde{Q}(-(\mathsf{I} - \hat{p}e')\hat{R}) + (-(\mathsf{I} - \hat{p}e')\hat{R})'\widetilde{Q}]\xi - 2\hat{p}'\widetilde{Q}(\mathsf{I} - \hat{p}e')\hat{R}\xi\phi(e'y))\\
&\quad +\kappa(1 - \dot{\phi}(e'y))(y'[-\widetilde{Q}\hat{R} - \hat{R}'\widetilde{Q}]y + \hat{\beta}y'\widetilde{Q}\hat{p} - \phi(e'y)\hat{p}'\widetilde{Q}\hat{R}y - \hat{p}'\widetilde{Q}\hat{p}\hat{\beta}).
\end{aligned} \tag{3.4.32}$$

Since $\xi'\hat{R}^{-1}\hat{p} = 0$, one checks as for (3.4.28) that there exists a constant $C > 0$ such that

$$\xi'[\widetilde{Q}(-(\mathsf{I} - \hat{p}e')\hat{R}) + (-(\mathsf{I} - pe')\hat{R})'\widetilde{Q}]\xi \leq -C|\xi|^2. \tag{3.4.33}$$

Moreover, from (3.4.22) and (3.4.31), we deduce that

$$y'[-\widetilde{Q}\hat{R} - \hat{R}'\widetilde{Q}]y \leq -C|y|^2 = -Ca^2 - C|\xi|^2. \tag{3.4.34}$$

Substituting (3.4.33) and (3.4.34) into (3.4.32), and using $0 \leq \dot{\phi}(e'y) \leq 1$ as well as $|\phi(e'y)| \leq \epsilon$, we obtain

$$(\nabla V(y))'b(y) \leq -2(a^2 + C|a||\xi| + C|a|) + \kappa(-C|\xi|^2 + C|\xi| + C|a| + C). \tag{3.4.35}$$

Since $e'y = a + e'\xi \in [-\epsilon, 0]$, we must have $|a| \leq C + |\xi|$ and consequently $|y| \leq C|a| + |\xi| \leq C|\xi| + C$. Thus for $|y|$ large, we can choose κ large so that the dominating term in (3.4.35) is $-\kappa C|\xi|^2$. Using the fact that $|y|^2 \leq C|\xi|^2$ when $|y|$ is large, we then deduce that there exists a constant $C > 0$ such that for $|y|$ large,

$$(\nabla V(y))' b(y) \leq -C|y|^2.$$

This concludes the proof for the third case.

On combining the above three cases we obtain that, for $|y|$ large,

$$(\nabla V(y))' b(y) \leq -C|y|^2,$$

as claimed in the proposition.

We now proceed to study the second derivative of V, which is denoted by $\ddot{V}$. We also write $\ddot{\phi}$ for the second derivative of ϕ. From (3.4.27), we find

$$\begin{aligned} \ddot{V}(y) &= 2ee' + 2\kappa[\widetilde{Q} + ee' \cdot \hat{p}'\widetilde{Q}\hat{p}(\ddot{\phi}(e'y)\phi(e'y) + \dot{\phi}(e'y)^2) \\ &\quad -(\widetilde{Q}pe' + e\hat{p}'\widetilde{Q})\dot{\phi}(e'y) - ee' \cdot y'\widetilde{Q}\hat{p}\ddot{\phi}(e'y)]. \end{aligned} \tag{3.4.36}$$

If $e'y \notin [-\epsilon, 0]$, we obtain $0 \leq \dot{\phi}(e'y) \leq 1$ and $\ddot{\phi}(e'y) = 0$. Therefore, for any i, j, there exists some $C > 0$ such that

$$\left| \frac{\partial^2 V}{\partial y_i \partial y_j}(y) \right| \leq C.$$

If $e'y \in [-\epsilon, 0]$, then $|\ddot{\phi}(e'y)| \leq C$ for some $C > 0$ since $\phi \in C^2(\mathbb{R})$ and $[-\epsilon, 0]$ is compact. Moreover, since $0 \leq \dot{\phi}(e'y) \leq 1$, the dominating term in (3.4.36) is $-2\kappa ee' \cdot y'\widetilde{Q}p\ddot{\phi}(e'y)$ for $|y|$ large. This implies that if $e'y \in [-\epsilon, 0]$ and $|y|$ is large, then there exists a constant $C > 0$ such that for any i, j,

$$\left| \frac{\partial^2 V}{\partial y_i \partial y_j}(y) \right| \leq C|y|,$$

where C is independent of y. This concludes the proof of (3.4.25). Now for $|y|$ large, we deduce from (3.4.25) that

$$GV(y) = (\nabla V(y))' b(y) + \frac{1}{2} \sum_{i,j} (\sigma\sigma')_{ij} \frac{\partial^2 V}{\partial y_i \partial y_j}(y) \leq -C|y|^2 \leq -1.$$

The proof of Proposition 3.4 is complete.

Proof of Theorem 3.3 In order to show that Y is positive recurrent and has a unique stationary distribution, we only have to check that $V(y) \to \infty$ as $|y| \to \infty$ in view of Proposition 3.1 and (3.4.26).

Let $x = e'y$ and $z = y - px^+$; then $|y|^2 \leq C(x^2 + |z|^2)$. We can rewrite (3.4.24) as follows:

$$\begin{aligned} V(y) &= x^2 + \kappa(y' - \hat{p}'\phi(x))\widetilde{Q}(y - \hat{p}\phi(x)) \\ &\geq x^2 + C|y - \hat{p}\phi(x)|^2 \\ &= x^2 + C|z + \hat{p}(x^+ - \phi(x))|^2 \\ &\geq x^2 + C|z|^2 - C\epsilon^2 \\ &\geq C|y|^2 - C\epsilon^2, \end{aligned}$$

where the second to last inequality uses the fact $0 \leq x^+ - \phi(x) \leq \frac{1}{2}\epsilon$. Therefore, $V(y) \to \infty$ as $|y| \to \infty$ and we conclude that Y has a unique stationary distribution.

To prove that Y is exponentially ergodic, we observe from (3.4.24) that there exists some $C > 0$ such that $V(y) \leq C|y|^2 + C$ for all $y \in \mathbb{R}^K$. Moreover, (3.4.26) implies that for $|y|$ large,

$$GV(y) \leq -CV(y) + C.$$

Putting this together with the fact that $V \in C^2(\mathbb{R}^K)$, we know that there exist some $c > 0$ and $d < \infty$ such that

$$GV(y) \leq -cV(y) + d \quad \text{for any } y \in \mathbb{R}^K.$$

Since $V \geq 0$, Proposition 3.2 implies that Y is f-exponentially ergodic, where $f = V + 1$. In particular, Y is exponentially ergodic since $f \geq 1$.

CHAPTER IV

INTERCHANGE OF LIMIT

4.1 Background

In this chapter we prove an interchange of limit theorem for $GI/Ph/n+M$ queues. $GI/Ph/n+M$ queues are a subset of $GI/GI/n+M$ queues. In a $GI/GI/n+M$ queue, the interarrival times of customers are i.i.d. following a general distribution (the first GI) and the service times are i.i.d. following a general distribution (the second GI). In addition, customers patience times are i.i.d. following an exponential distribution ($+M$).

As mentioned in the introduction, despite of decades of research, there is still no general analytical or numerical tool to efficiently and accurately predict the steady-state performance of a $GI/GI/n+M$ system. Computer simulation is often the only remaining tool available, but it can be slow when the number of servers is large, particularly when the system is in the QED regime.

Dai and He [25] recently proposed diffusion models to approximate a $GI/Ph/n+M$ system when the service time distribution is of phase-type (see Section 2.1.1 for a definition). The numerical examples in [25] demonstrate that the steady-state performance of the diffusion model provides a remarkably accurate estimate for the steady-state performance of the corresponding queueing system, even when the number of servers is moderate. In this chapter, we provide a justification for the diffusion approximation procedure in [25].

The rest of this chapter is organized as follows. Section 4.2 presents the assumptions on $GI/Ph/n+M$ queues. Assuming positive Harris recurrence, Section 4.3 summarizes our main results, Theorem 4.1 and Corollary 4.1. Section 4.4 introduces

our Lyapunov function and a fluid model used to prove Lemma 4.3. Section 4.5 discusses a key lemma, Lemma 4.3, for proving Theorem 4.1. In Appendix A.3, we prove the positive Harris recurrence of $GI/Ph/n+M$ queues when n is fixed. The proof of a negative drift condition for the fluid model is given in Appendix A.4. Appendix A.5 uses this negative drift condition for the fluid model to establish a negative drift condition for the diffusion-scaled processes.

4.2 Assumptions

We discuss the assumptions on $GI/Ph/n+M$ queues in this section. Since $GI/Ph/n+M$ queues are a subset of $G/Ph/n+GI$ queues introduced in Chapter 2, we use the same notation for the state processes and our presentation focuses on their difference.

In a $GI/Ph/n+M$ queue, customers arrive according to a delayed renewal process with interarrival times given by $\{\xi^n(i) : i = 0, 1, 2, \ldots\}$. We assume that $\{\xi^n(i) : i = 1, 2, \ldots\}$ is a sequence of i.i.d. random variables with a general distribution and this sequence is independent of $\xi^n(0)$. Here $\xi^n(0)$ is the time that the first customer after time 0 is to arrive at the system. Service times are i.i.d. following a phase-type distribution. In addition, the patience times of customers follow an pexponential distribution with positive rate $\alpha > 0$. We assume the sequence of $GI/Ph/n+M$ queues is in the QED regime introduced in Subsection 2.2.1.

Given a $GI/Ph/n+M$ queue for fixed n, recall from Section 2.2.2 that $Z^n = \{Z^n(t) : t \geq 0\}$ is the server-allocation process and $N^n(t)$ is the number of customers in the n-th system at time t, either in queue or in service. Let $A^n(t)$ be the "age" of the interarrival time at time t, i.e., the time between the arrival time of the most recent arrival before time t and time t. Then, for fixed n, (A^n, N^n, Z^n) has state space $\mathcal{S} = \mathbb{R}_+ \times \mathbb{Z}_+ \times \mathbb{Z}_+^K$. As time t goes on, $A^n(t)$ increases linearly while $(N^n(t), Z^n(t))$ remains constant. When $A^n(t)$ reaches the interarrival for the next arrival, it instantaneous jumps to zero. We adopt the convention that all processes are right continuous

on $[0,\infty)$, having left limits in $(0,\infty)$. It follows that (A^n, N^n, Z^n) is a piecewise deterministic Markov process that conforms to Assumption 3.1 of Davis [28], and hence it is a strong Markov process Davis [28, page 362]. Throughout this chapter, we make the following assumptions.

Assumption 4.1 *The interarrival times* $\{\xi^n(i) : i = 1, 2, \ldots\}$ *from the second customer onwards have the following representation:* $\xi^n(i) = \frac{1}{\lambda^n} u(i)$, $i \geq 1$, *where* $\{u(i) : i \geq 1\}$ *is an i.i.d. sequence with* $\mathbb{E}(u(i)) = 1$ *and* $\mathbb{E}(u(i))^2 < \infty$.

Assumption 4.2 *For each* n, *the Markov process* (A^n, N^n, Z^n) *is positive Harris recurrent and has stationary distribution* π^n.

For the definition of positive Harris recurrence of a Markov process, see, for example, Dai [22]. Proposition A.1 in Appendix A.3 provides a sufficient condition on the interarrival distribution for Assumption 4.2 to hold.

To facilitate the analysis, we define the diffusion-scaled age process:

$$\tilde{A}^n(t) = \frac{1}{\sqrt{n}} A^n(t). \tag{4.2.1}$$

Recall $\tilde{X}^n$ and $\tilde{Z}^n$ defined in (2.2.6) and (2.2.5). We write $\tilde{\mathcal{S}}^n$ for the state space of $(\tilde{A}^n, \tilde{X}^n, \tilde{Z}^n)$:

$$\tilde{\mathcal{S}}^n = \{(a, x, z) \in \mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}^K : \sqrt{n}x + n \in \mathbb{Z}_+, \sqrt{n}z + n\gamma \in \mathbb{Z}_+^K, e'z + x^- = 0\}.$$

Using Assumption 4.1 on the interarrival times, we deduce from the functional central limit theorem that (2.2.2) holds. Thus Theorem 2.1 holds. In particular,

$$(\tilde{X}^n, \tilde{Z}^n) \Rightarrow (\tilde{X}, \tilde{Z}),$$

where $(\tilde{X}, \tilde{Z})$ is a Markov process living on

$$\tilde{\mathcal{S}} = \{(x, z) \in \mathbb{R} \times \mathbb{R}^K : e'z + x^- = 0\}. \tag{4.2.2}$$

Since we have assumed the abandonment rate $\alpha > 0$, we can apply Theorem 3.3. We therefore obtain from (2.2.13) that the Markov process $(\tilde{X}, \tilde{Z})$ is positive recurrent and has a unique stationary distribution π.

4.3 Result on interchange limit

In this section, we state our main result in this chapter and work under Assumption 4.2. Let $\tilde{\pi}_n$ be the stationary distributions of the diffusion-scaled state process $(\tilde{A}^n, \tilde{X}^n, \tilde{Z}^n)$ defined in (4.2.1), (2.2.6) and (2.2.5). Since $\tilde{A}^n \Rightarrow 0$ as $n \to \infty$, we focus on the marginal distributions of $\{(\tilde{X}^n, \tilde{Z}^n) : n \geq 1\}$. Now we state the main theorem of this chapter. See Billingsley [11] for concepts and details on tightness.

Theorem 4.1 *Suppose that Assumptions 4.1 and 4.2 and the many-server heavy traffic condition (2.2.1) hold. Assume that the abandonment rate α is strictly positive. Let $(\tilde{X}^n(0), \tilde{Z}^n(0))$ be distributed according to the stationary distribution $\tilde{\pi}^n$. Then the sequence of random vectors $\{(\tilde{X}^n(0), \tilde{Z}^n(0)) : n \geq 1\}$ is tight.*

The following corollary states the validity of interchange of heavy-traffic and steady-state limits. Because $(\tilde{X}, \tilde{Z})$ has a unique stationary distribution (see Theorem 3.3 and Equation (2.2.13)), the corollary follows from Theorem 4.1 by a standard argument; see Gamarnik and Zeevi [42] and Budhiraja and Lee [16].

Corollary 4.1 *Under the assumptions of Theorem 4.1, the sequence of marginal distributions of $\tilde{\pi}^n$ on $(\tilde{X}^n(0), \tilde{Z}^n(0))$ converges weakly to π, where π is the unique stationary distribution of $(\tilde{X}, \tilde{Z})$ and $(\tilde{X}, \tilde{Z})$ is the Markov process from Theorem 2.1.*

4.4 Our Lyapunov function and a fluid model

In this section, we first introduce the Lyapunov function that lies at the heart of this work. We then introduce a fluid model associated with the sequence of $GI/Ph/n+M$ systems, and assert that our Lyapunov function is a geometric Lyapunov function for the fluid model. The proof of our main result relies on a comparison between the diffusion-scaled processes and the fluid model.

To define our Lyapunov function, we use common quadratic Lyapunov functions in Theorem 3.1. Recall that p is a discrete probability distribution on $\{1, \ldots, K\}$,

representing the initial distribution of the phase type service time distribution, e is a vector of ones, and R is a matrix given in (2.2.8) and is obtained from the parameters of the phase-type service time distribution. As discuss at the end of Chapter 2, R takes the form of $(\mathsf{I} - P')\,\mathrm{diag}\{\nu\}$, where P is a transient matrix. We thus deduce R is a nonsingular M-matrix and $e'R \geq 0$. In particular, Assumption 2.1 is satisfied. It follows from Theorem 3.1 that there exists a $K \times K$ positive definite matrix $\tilde{Q}$ such that

$$\tilde{Q}R + R'\tilde{Q} \quad \text{is positive definite,} \tag{4.4.1}$$

$$\tilde{Q}(\mathsf{I} - pe')R + R'(\mathsf{I} - ep')\tilde{Q} \quad \text{is positive semi-definite.} \tag{4.4.2}$$

Our Lyapunov function is the square root of a function $g = g_\beta$ that depends on the "slack parameter" β in (2.2.1). It is defined as

$$g(x,z) = \begin{cases} (x+\beta)^2 + \kappa(z+\beta\gamma)'\tilde{Q}(z+\beta\gamma), & \text{if} \quad \beta \geq 0, \\ (\alpha x + \mu\beta)^2 + \kappa(z+\beta\gamma)'\tilde{Q}(z+\beta\gamma), & \text{if} \quad \beta < 0, \end{cases} \tag{4.4.3}$$

where γ is defined in (5.5.2) and κ is a large positive constant to be determined later.

The quadratic function in (4.4.3) is slightly different from the Lyapunov function used in Chapter 3. The modification is needed because Chapter 3 focused on a K-dimensional state process $\tilde{Y}$, but here we focus on the degenerate $(K+1)$-dimensional state process $(\tilde{X}, \tilde{Z})$ on the manifold $\tilde{\mathcal{S}}$; both state processes are equivalent. More importantly, the centering for (x, z) in the two quadratic terms is different. Our centering allows us to obtain a stronger property for the fluid model; see part (a) of Lemma A.7.

We now introduce a fluid model associated with $GI/Ph/n + M$ systems. This fluid model is defined through a map, which is established in a more general setting in Dai et al. [24]. Recall the definition of $\tilde{\mathcal{S}}$ in (4.2.2) and write

$$\tilde{\mathcal{T}} = \{(u, v) \in \mathbb{R} \times \mathbb{R}^K : e'v = 0\}.$$

Lemma 4.1 *Let $\alpha > 0$.*

(a) For each $(u, v) \in \mathbb{D}(\tilde{\mathcal{T}})$, there exists a unique $(x, z) \in \mathbb{D}(\tilde{\mathcal{S}})$ such that

$$x(t) = u(t) - \alpha \int_0^t (x(s))^+ \, ds - e'R \int_0^t z(s) \, ds, \tag{4.4.4}$$

$$z(t) = v(t) - p(x(t))^- - (\mathsf{I} - pe')R \int_0^t z(s) \, ds \tag{4.4.5}$$

for $t \geq 0$.

(b) For each $(u, v) \in \mathbb{D}(\tilde{\mathcal{T}})$, define $(x, z) = \Psi(u, v) \in \mathbb{D}(\tilde{\mathcal{S}})$, where (x, z) satisfies (4.4.4) and (4.4.5). The map Ψ is well-defined and is continuous when both the domain $\mathbb{D}(\tilde{\mathcal{T}})$ and the range $\mathbb{D}(\tilde{\mathcal{S}})$ are endowed with the standard Skorohod J_1-topology.

(c) The map Ψ is Lipschitz continuous in the sense that for any $T > 0$, there exists a constant $C = C(T) > 0$ such that

$$\|\Psi(y^1) - \Psi(y^2)\|_T \leq C\|y^1 - y^2\|_T \quad \text{for any } y^1, y^2 \in \mathbb{D}(\tilde{\mathcal{T}}).$$

(d) The map Ψ is positively homogeneous in the sense that

$$\Psi(by) = b\Psi(y) \quad \text{for each } b > 0 \text{ and each } y \in \mathbb{D}(\tilde{\mathcal{T}}).$$

We now define the fluid counterpart of the diffusion-scaled state processes $(\tilde{X}^n, \tilde{Z}^n)$. Fix an initial state $(\tilde{x}(0), \tilde{z}(0)) \in \tilde{\mathcal{S}}$. Set

$$\tilde{u}(t) = \tilde{x}(0) - \mu\beta t \quad \text{and} \quad \tilde{v}(t) = (\mathsf{I} - pe')\tilde{z}(0), \tag{4.4.6}$$

and, after noting that $(\tilde{u}(0), \tilde{v}(0)) \in \mathbb{D}(\tilde{\mathcal{T}})$, set

$$(\tilde{x}, \tilde{z}) = \Psi(\tilde{u}, \tilde{v}). \tag{4.4.7}$$

We call $(\tilde{x}, \tilde{z})$ the fluid model starting from $(\tilde{x}(0), \tilde{z}(0))$. The next lemma is a negative drift condition for the fluid model, and states that the function $\sqrt{g}$ is a geometric Lyapunov function for the fluid model. Appendix A.4 is devoted to its proof.

Lemma 4.2 *Fix some $t_0 > 0$. There exists constants $C = C(t_0) > 0$ and $\epsilon = \epsilon(t_0) \in (0,1)$ such that for each initial state $(x,z) = (\tilde{x}(0), \tilde{z}(0)) \in \tilde{\mathcal{S}}$, we have*

$$\sqrt{g(\tilde{x}(t_0), \tilde{z}(t_0))} - \sqrt{g(x,z)} \leq C - \epsilon\sqrt{g(x,z)}.$$

4.5 Our Lyapunov function and diffusion-scaled processes

In this section, we present a negative drift condition for the diffusion-scaled processes, and we briefly outline how this leads to our main result. We use the same Lyapunov function $\sqrt{g}$ as in the fluid model.

We are now ready to formulate our negative drift condition for diffusion-scaled processes. Here and in the rest of this chapter, we adopt the notational convention that

$$\mathbb{E}_{(a,x,z)}[\quad \cdot \quad] = \mathbb{E}[\quad \cdot \quad | \tilde{A}^n(0) = a, \tilde{X}^n(0) = x, \tilde{Z}^n(0) = z]$$

for each initial state $(a,x,z) \in \tilde{\mathcal{S}}^n$.

Lemma 4.3 *Fix any $t_0 > 0$. Under the assumptions of Theorem 4.1, there exists a nonnegative function g on $\tilde{\mathcal{S}}$, as well as two constants $C = C(t_0) > 0$ and $\epsilon = \epsilon(t_0) \in (0,1)$, such that for each n and each feasible initial state $(a,x,z) \in \tilde{\mathcal{S}}^n$, we have*

$$\mathbb{E}_{(a,x,z)}\left[\sqrt{g(\tilde{X}^n(t_0), \tilde{Z}^n(t_0))}\right] - \sqrt{g(x,z)} \leq C - \epsilon\sqrt{g(x,z)}. \tag{4.5.1}$$

The function $\sqrt{g}$ satisfying (4.5.1) is essentially a geometric Lyapunov function with a geometric drift size $1-\epsilon$ and drift time t_0, see for instance Section 3 in Gamarnik and Zeevi [42]. Readers are referred to Gamarnik and Zeevi [42] and Meyn and Tweedie [76] for more details on the definition of a Lyapunov function and its application in deriving bounds for stationary distributions of Markov processes.

The proof of the negative drift condition in Lemma 4.3 is lengthy and will be given in Appendix A.5. Assuming the lemma, the proof of Theorem 4.1 is standard. We end this section by giving a sketch of the proof, which is almost identical to the proof of Theorem 5 in Gamarnik and Zeevi [42].

Proof sketch of Theorem 4.1 In order to show that $\tilde{\pi}^n$ is tight, since $\tilde{A}^n \Rightarrow 0$ as $n \to \infty$ and g has compact level sets (see Appendix A.4), it is sufficient to show that for any given $\delta > 0$, there exists some large constant s, such that for all n sufficiently large,

$$P_{\tilde{\pi}^n}(\sqrt{g(\tilde{X}^n(0), \tilde{Z}^n(0))} > s) \leq \delta.$$

By Markov's inequality, it suffices to show that, for some $C < \infty$,

$$\mathbb{E}_{\tilde{\pi}^n}\left[\sqrt{g(\tilde{X}^n(0), \tilde{Z}^n(0))}\right] \leq C/\epsilon, \tag{4.5.2}$$

where C and ϵ are constants in (4.5.1). To prove (4.5.2), we use (4.5.1) in the following form:

$$\epsilon\sqrt{g(x,z)} - C \leq \sqrt{g(x,z)} - \mathbb{E}_{(a,x,z)}[\sqrt{g(\tilde{X}^n(t_0), \tilde{Z}^n(t_0))}].$$

We argue that the right-hand side is nonpositive after taking the expectation with respect to $\tilde{\pi}^n$, and (4.5.2) then follows. For each integer $k \geq 1$ and each $(x, z) \in \mathbb{R} \times \mathbb{R}^K$, set $f_k(x,z) = \sqrt{g(x,z)} \wedge k$. It can be checked that $f_k(x,z) - \mathbb{E}_{(a,x,z)} f_k(\tilde{X}^n(t_0), \tilde{Z}^n(t_0))$ is bounded below by $-C$ for all $(a, x, z) \in \tilde{\mathcal{S}}^n$. Therefore, Fatou's Lemma can be applied and we deduce that

$$\begin{aligned}
&\int_{\tilde{\mathcal{S}}} \left[\sqrt{g(x,z)} - \mathbb{E}_{(a,x,z)}\sqrt{g(\tilde{X}^n(t_0), \tilde{Z}^n(t_0))}\right] d\tilde{\pi}^n(a,x,z) \\
&\quad \leq \liminf_{k\to\infty} \int_{\tilde{\mathcal{S}}} \left[f_k(x,z) - \mathbb{E}_{(a,x,z)} f_k(\tilde{X}^n(t_0), \tilde{Z}^n(t_0))\right] d\tilde{\pi}^n(a,x,z) = 0,
\end{aligned}$$

where the equality follows from stationarity of $\tilde{\pi}^n$.

STOCHASTIC MODELS FOR SERVICE SYSTEMS AND LIMIT ORDER BOOKS

PART II

Stochastic networks

by

Xuefeng Gao

CHAPTER V

SENSITIVITY ANALYSIS OF DIFFUSION PROCESSES CONSTRAINED TO AN ORTHANT

5.1 Introduction

The second part of the thesis is on sensitivity analysis in stochastic networks. We are motivated by a desire to better understand the relation between performance metrics and control variables in a network with shared but limited resources. We are specifically interested in service networks, where customers seeking a certain service may suffer from delays as a result of temporary insufficient service capacity. The control variables are the service capacities at the individual stations. Many service processes can be modeled by stochastic (or queueing) networks, and an important question is how resources should be allocated given random fluctuations in the arrivals and its interplay with potentially random service times. When planning horizons are long so that static allocation rules are required, questions of this type are readily answered if the network has a product-form structure Kleinrock [65], Wein [105]. However, few results have been obtained when this assumption fails (Dieker et al. [32], Pollett [83]). It is the goal of this chapter to introduce new tools in this context, which could be used in the context of both sensitivity analysis and system optimization.

We study diffusion processes and their 'derivatives', defined as the change in the process under an infinitesimal change in the drift. Although some of our results are stated more generally, this chapter focuses on diffusion processes for two reasons. First, this framework allows us to explain key concepts in a tractable yet relatively general setting. Second, diffusion processes are rooted in heavy-traffic approximations for stochastic networks, and the heavy-traffic assumption seems reasonable in the

context of resource allocation problems with systems operating close to their capacity. This chapter studies the stationary distribution of diffusions and their derivatives, as a proxy for the long-term (steady-state) behavior. Although it is certainly desirable to obtain time-dependent tools as well, given the vast body of work on stationary results, making this assumption is a natural first step. The techniques developed in this chapter are likely to be also relevant in the time-dependent case.

We have two main results in this chapter. The first is a statement on the behavior of deterministic functions under the well-known Skorohod reflection map with oblique reflection (regulation), and states that the map and its 'derivative' are the unique solution to an augmented version of the Skorohod problem. Our proof of this result relies on recent insights into directional derivatives by Mandelbaum and Ramanan [74], which have been developed in the context of time-inhomogeneous systems but are shown here to be useful for sensitivity analysis as well.

Our second main result specializes to diffusion processes and studies the stationary distribution of solutions to the augmented Skorohod problem. Given a constrained diffusion process Z representing the dynamics of the underlying stochastic network (i.e., the queue lengths at each of the stations), let the stochastic process A represent the change in Z under an infinitesimal change in the drift. The two results combined say that the stationary distribution of the joint processes (Z, A) satisfies a kind of basic adjoint relation, which is the analog of the equation $\pi' Q = 0$ for continuous-time Markov processes on a discrete state space. The proof relies on a delicate analysis of the jumps of A; the process A has jumps even if Z is continuous.

The intuition behind the program carried out in this chapter can be summarized as follows. Suppose Z^ϵ is a constrained diffusion process with drift coefficient $\mu(\cdot) - \epsilon v$ in the interior of the orthant, where v is an arbitrary nonnegative vector. Suppose the processes $\{\mathsf{Z}^\epsilon\}$ are driven by the same Brownian motion for every $\epsilon \geq 0$, so that they are coupled. The processes $\mathsf{Z} \equiv \mathsf{Z}^0$ and Z^ϵ are Markovian, and one can therefore

expect to be able to give a basic adjoint relationship for their stationary distributions (should they exist). Moreover, $(\mathsf{Z}, \mathsf{Z}^\epsilon)$ and therefore $(\mathsf{Z}, (\mathsf{Z} - \mathsf{Z}^\epsilon)/\epsilon)$ can be expected to be Markovian as a result of the coupling. Provided one can make sense of the pointwise limit (Z, A) of $(\mathsf{Z}, (\mathsf{Z}-\mathsf{Z}^\epsilon)/\epsilon)$ as $\epsilon \to 0+$, one can expect that the distribution of (Z, A) satisfies a similar relationship. This results in an 'augmented' basic adjoint relationship, which we state in Theorem 5.3. The constrained diffusion processes studied in this chapter are pathwise solutions to stochastic differential equations with reflection, see Dupuis and Ishii [35], Ramanan [88]. We only consider left derivatives in this work, although one could develop similar tools and obtain similar results for right derivatives. This would affect our two main results as follows. On a sample-path level, the right derivative is the left-continuous modification of the (right-continuous) left derivative, see Section 5.4.1 for a detailed discussion. On a probabilistic level, studying the (left-continuous) right derivative requires a different set of technical tools since one ordinarily works with right-continuous stochastic processes. We should expect that this change does not affect the stationary distribution or the basic adjoint relationship.

When carrying out the aforementioned approach, we were surprised to find that, even though Z is known not to spend any time on low-dimensional faces, it is critical to incorporate the jumps of A when Z reaches those faces in order to formulate the basic adjoint relationship.

This work has the potential to lead to new numerical methods in the context of optimization and sensitivity analysis for queueing networks, which relieve or remove the need for computationally intensive or numerically unstable operations such as gradient estimation. To explain, due to the division by ϵ, any performance metric of $(\mathsf{Z} - \mathsf{Z}^\epsilon)/\epsilon$ suffers from numerical instability issues for small $\epsilon > 0$. Researchers in stochastic optimization have developed several techniques to mitigate this effect (see, e.g., Asmussen and Glynn [7]). The approach taken in this work is to analytically

describe and investigate the dynamics of the limit. Our experience with state-of-the-art stochastic optimization implementations in the context of resource capacity management, as documented in part in Dieker et al. [32], is that it is computationally very costly to obtain reliable gradient estimates and that the use of 'quick and dirty' estimates can have disastrous effects on the compute time of a stochastic optimization procedure due to bias and inherent random fluctuations. Therefore, reliable (numerical) tools that give merely a rough idea of the gradient can be desirable and useful. In particular, from an implementation perspective, heavy-traffic gradient information can be valuable even if a stochastic network is in moderate traffic. (A light-traffic setting is not of prime interest since one is typically interested in fine-tuning networks operating in a regime where servers are idling relatively rarely.)

The framework of this work is related to a body of literature known as infinitesimal perturbation analysis (Glasserman [44], Glasserman [46], Glasserman [45], Heidergott [53]). Infinite perturbation analysis also aims to perform sensitivity analysis or gradient estimation for performance metrics in (say) a queueing network, and it does so by formulating conditions under which an expectation and a derivative operator can be interchanged. Here, however, it is not our objective to seek such an interchange involving a performance metric but instead we study the (whole) stationary distribution of a stochastic process with its derivative process.

The rest of this chapter is outlined as follows. Section 5.2 summarizes our approach in the one-dimensional case, which serves as a guide for our multi-dimensional results. Section 5.3 discusses two technical preliminaries: oblique reflection maps and their derivatives. In Section 5.4 we formulate our two main results. Section 5.5 is devoted to the proof of the first main result, while Section 5.6 gives the proof of the second main result. A key role is played by jump measures, for which we obtain a description in Section 5.7. Several technical digressions are given in appendices.

5.1.1 Notation

This subsection contains all the notations used in this chapter. For $J \in \mathbb{N}$, $\mathbb{R}^J$ denotes the J-dimensional Euclidean space. We denote the space of real $n \times m$ matrices by $\mathbb{M}^{n\times m}$, and the subset of nonnegative matrices by $\mathbb{M}_+^{n\times m}$. All vectors are to be interpreted as column vectors, and we write M^j and M_i for the j-th column and the i-th row of a matrix M, respectively. In particular, v_i is the i-th element of a vector v and M_i^j is element (i, j) of a matrix M. Similarly, given a set $I \subseteq \{1, \ldots, J\}$, we write M_I and M^I for the matrices consisting of the rows and columns of M, respectively, with indices in I. Throughout, I stands for the identity matrix and we write δ_i^j for I_i^j. We use the symbol $'$ for transpose. The norms $\|\cdot\|_1$ and $\|\cdot\|_2$ stand for entrywise 1-norm and 2-norm, respectively, and is used for both vectors and matrices.

Given a measure space $(S, \mathcal{S})$, a measurable vector-valued function $h : S \to \mathbb{R}^J$ on $(S, \mathcal{S})$, and a vector of measures $\nu = (\nu_1, \ldots \nu_J)$ on $(S, \mathcal{S})$, we set

$$\int h(x)\nu(dx) = \int h(x) \cdot \nu(dx),$$

provided the right-hand side exists. We shall also employ this notation when h and ν are matrix-valued. That is, we write for $h : S \to \mathbb{M}^{J\times J}$ and an $\mathbb{M}^{J\times J}$-valued measure ν on $(S, \mathcal{S})$,

$$\int h(x)\nu(dx) = \int \langle h(x), \nu(dx)\rangle_{\text{HS}},$$

where $\langle\cdot,\cdot\rangle_{\text{HS}}$ is the Hilbert-Schmidt inner product on $\mathbb{M}^{J\times J}$ given by

$$\langle M_1, M_2\rangle_{\text{HS}} = \text{tr}(M_1' M_2).$$

For a function $g : \mathbb{M}^{J\times J} \to \mathbb{R}$, we define $\nabla g : \mathbb{M}^{J\times J} \to \mathbb{M}^{J\times J}$ as the function for which element (i, j) is given by the directional derivative of g in the direction of the matrix with only zero entries except for element (i, j), where its entry is 1. We also write, for $i = 1, \ldots, J$, $F_i = \{(\mathsf{z}, a) \in \mathbb{R}_+^J \times \mathbb{M}^{J\times J} : z_i = 0\}$, $F_i^a = \{(z, a) \in \mathbb{R}_+^J \times \mathbb{M}^{J\times J} : a_i = 0\}$. The space of functions $f : \mathbb{R}_+^J \times \mathbb{M}_+^{J\times J} \to \mathbb{R}$ which are twice continuously differentiable with bounded derivatives is denoted by $C_b^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$.

We write $\mathbb{D}_+^J$ for the space of $\mathbb{R}_+^J$-valued functions on $\mathbb{R}_+$ which are right-continuous on $\mathbb{R}_+$ with left limits in $(0,\infty)$. The subset of continuous functions is written as C^J, and C_+^J denotes the set of nonnegative continuous functions. Similarly, we write $\mathbb{D}^{J\times J}$ for the space of $\mathbb{M}^{J\times J}$-valued right-continuous functions on $\mathbb{R}_+$ with left limits. The subset of $\mathbb{M}_+^{J\times J}$-valued functions is denoted by $\mathbb{D}_+^{J\times J}$.

5.2 A motivating one-dimensional result

Fix some $\theta < 0$. For any $\epsilon \geq 0$, we let Z^ϵ be a one-dimensional reflected Brownian motion with drift $\theta - \epsilon < 0$ and variance σ^2. That is,

$$\mathsf{Z}^\epsilon(t) = \mathsf{X}^\epsilon(t) + \mathsf{Y}^\epsilon(t) \geq 0,$$

where X^ϵ is a Brownian motion with drift $\theta - \epsilon$ and variance σ^2, and the regulating term Y^ϵ is given by

$$\mathsf{Y}^\epsilon(t) = \max\left(\sup_{0\leq s\leq t}[-\mathsf{X}^\epsilon(s)], 0\right).$$

Suppose the family $\{\mathsf{Z}^\epsilon : \epsilon \geq 0\}$ is coupled in the sense that $\mathsf{X}^\epsilon(t) = W(t) + (\theta - \epsilon)t$ for some driftless Brownian motion W. Write $\mathsf{Z} \equiv \mathsf{Z}^0$.

It follows from Theorem 1.1 in Mandelbaum and Ramanan [74] (see also Lemma 5.2 and Equation (5.7) in Mandelbaum and Massey [73]) that, for each $t \geq 0$, the limit

$$\mathsf{A}(t) \equiv \lim_{\epsilon\to 0+} \frac{1}{\epsilon}(\mathsf{Z}(t) - \mathsf{Z}^\epsilon(t)) \tag{5.2.1}$$

exists. We also have the following explicit formula:

$$\mathsf{A}(t) = t - \mathsf{B}(t), \tag{5.2.2}$$

where

$$\mathsf{B}(t) = \sup\{s \in [0,t] : \mathsf{Z}(s) = 0\},$$

and $\sup\emptyset = 0$ by convention. In view of the definition of A in (5.2.1), we call it the derivative process of Z.

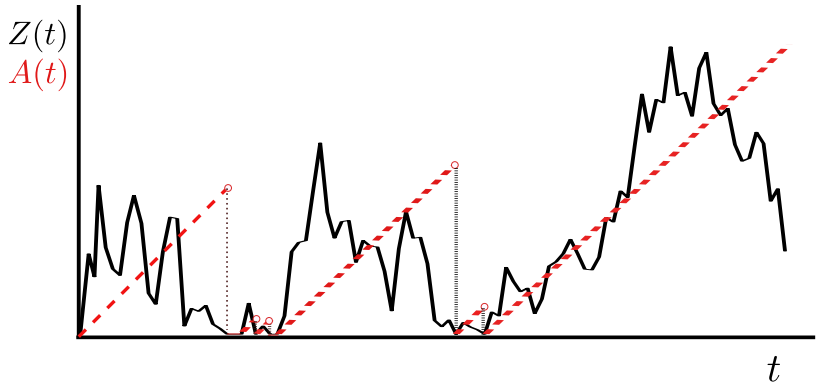

Figure 1: Sample paths of (Z, A) as a function of time. The solid black curve is Z, while the dashed red curve is A. The slope of A is 1 whenever it is continuous, and A jumps to 0 whenever Z hits 0.

We now relate these notions to sensitivity analysis. Our investigations are motivated by the following sequence of equalities: for any 'smooth' function (performance measure) ϕ, one could expect that

$$\frac{d}{d\epsilon}\mathbb{E}\left[\phi(\mathsf{Z}^\epsilon(\infty))\right] = \mathbb{E}\left[\frac{d}{d\epsilon}\phi(\mathsf{Z}^\epsilon(\infty))\right] = \mathbb{E}\left[\mathsf{A}(\infty)\phi'(\mathsf{Z}(\infty))\right]. \tag{5.2.3}$$

Thus, to study (infinitesimal) changes in the steady-state performance measure under infinitesimal changes in the drift θ, one is led to investigating the stationary distribution of (Z, A) (assuming it exists). We are able to justify the interchange of expectation and derivative in the above equalities in the one-dimensional case (see below), but a justification in the setting of general multidimensional constrained diffusions requires a different set of techniques and falls outside the scope of this work.

One readily checks that the sample paths of the process B are nondecreasing, that they are right-continuous with left-hand limits, and that A has positive drift and negative jumps. In particular, the process A is of finite variation and (Z, A) is a semimartingale with jumps. An illustration of the process (Z, A) is given in Figure 1. From Ito's formula in conjunction with sample path properties of A, we obtain the following result. We suppress further details of the proof, since this program is carried out in greater generality in Section 5.6.

Theorem 5.1 *Let* Z *be a one-dimensional reflected Brownian motion with drift* θ *and variance* σ^2*. Let* A *be defined in (5.2.2). Suppose that the process* (Z, A) *has a unique stationary distribution* π*. For any* $f \in C_b^2(\mathbb{R}_+ \times \mathbb{R}_+)$*, we have the following relationship:*

$$\begin{aligned} 0 &= \int_0^\infty \int_0^\infty \left[\frac{1}{2}\sigma^2 \frac{\partial^2}{\partial \mathsf{z}^2} f(\mathsf{z},\mathsf{a}) + \theta \frac{\partial}{\partial \mathsf{z}} f(\mathsf{z},\mathsf{a}) + \frac{\partial}{\partial \mathsf{a}} f(\mathsf{z},\mathsf{a}) - \frac{\partial}{\partial \mathsf{a}} f(0,\mathsf{a})\right] \pi(d\mathsf{z}, d\mathsf{a}) \\ &\quad - \frac{\partial}{\partial \mathsf{z}} f(0,0)\theta. \end{aligned} \tag{5.2.4}$$

One can go further and derive the Laplace transform of π using this theorem, see Appendix B.1. One then finds that, for any $\alpha, \eta > 0$,

$$\int_0^\infty \int_0^\infty e^{-\alpha \mathsf{z} - \eta \mathsf{a}} \pi(d\mathsf{z}, d\mathsf{a}) = \frac{-2\theta}{\alpha\sigma^2 - \theta + \sqrt{2\eta\sigma^2 + \theta^2}}. \tag{5.2.5}$$

In particular, the theorem completely determines the stationary measure π. It is also possible to derive this result immediately from standard fluctuation identities for Brownian motion with drift, using results from Dębicki et al. [30]. In fact, since the corresponding densities are known explicitly (or can be found by inverting the Laplace transform), it is possible to write down the density of $(\mathsf{Z}(\infty), \mathsf{A}(\infty))$ in closed form. Using the resulting expression, it can be verified directly that (5.2.3) indeed holds.

5.3 Oblique reflection maps and their directional derivatives

This section contains the technical preliminaries to formulate a multidimensional analog of Theorem 5.1. We need the following definition to introduce the analogs of the processes A and B.

Definition 5.1 *(Oblique reflection map) Suppose a given* $J \times J$ *real matrix* R *can be written as* $\mathsf{R} = \mathsf{I} - \mathsf{P}$*, where* P *is a nonnegative matrix with spectral radius less than one and zeros on the diagonal. Then for every* $\mathsf{x} \in \mathbb{D}^J$*, there exists a unique pair* $(\mathsf{y}, \mathsf{z}) \in \mathbb{D}_+^J \times \mathbb{D}_+^J$ *satisfying the following conditions:*

1. $\mathsf{z}(t) = \mathsf{x}(t) + \mathsf{R}\mathsf{y}(t) \geq 0$ *for* $t \geq 0$,

2. $\mathsf{y}(0) = 0$, y *is componentwise nondecreasing and*

$$\int_0^\infty \mathsf{z}(t) d\mathsf{y}(t) = 0.$$

We write $\mathsf{y} = \Phi(\mathsf{x})$ *and* $\mathsf{z} = \Gamma(\mathsf{x})$ *for the oblique reflection map.*

The reflection map gives rise to left derivatives as formalized in the following definition. Existence of the derivatives is guaranteed by Theorem 1.1 in Mandelbaum and Ramanan [74].

Definition 5.2 *(Derivatives of the reflection map) Let* $\chi(t) = t\mathsf{I}$ *and define the* $\mathbb{M}^{J\times J}$*-valued functions* a *and* b *by defining* $\mathsf{a} = \lim_{\epsilon\to 0+} \mathsf{a}_\epsilon$ *and* $\mathsf{b} = \lim_{\epsilon\to 0+} \mathsf{b}_\epsilon$*, where the limits are to be understood as pointwise limits and, for* $j = 1, \ldots, J$*,*

$$\begin{aligned} \mathsf{a}_\epsilon^j &\equiv \frac{1}{\varepsilon}\left[\Gamma(\mathsf{x}) - \Gamma(\mathsf{x} - \epsilon\chi^j)\right], \\ \mathsf{b}_\epsilon^j &\equiv -\frac{1}{\varepsilon}\left[\Phi(\mathsf{x}) - \Phi(\mathsf{x} - \epsilon\chi^j)\right]. \end{aligned} \tag{5.3.1}$$

Then we have for each $t \geq 0$*,*

$$\mathsf{a}(t) = t\mathsf{I} - \mathsf{R}\mathsf{b}(t). \tag{5.3.2}$$

For notational convenience, we write $\mathsf{a} = \Gamma'(\mathsf{x})$ *and* $\mathsf{b} = -\Phi'(\mathsf{x})$*.*

5.4 Main results

This section states the main results of this chapter. The first result makes the connection between derivatives and an augmented Skorohod problem, which we define momentarily. The second result is a basic adjoint relationship for the stationary distribution of solutions to the augmented Skorohod problem with diffusion input. The basic adjoint relationship is the analog of the equation $\pi' Q = 0$ for Markov chains on a countable state space as mentioned in the introduction.

5.4.1 Augmented Skorohod problems and derivatives

In this section we introduce the augmented Skorohod problem and connect it with derivatives of the oblique reflection map.

Definition 5.3 *(Augmented Skorohod problem) Suppose we are given two $J \times J$ real matrices $\mathsf{R} = \mathsf{I} - \mathsf{P}$ and $\tilde{\mathsf{R}} = \mathsf{I} - \tilde{\mathsf{P}}$, where both P and $\tilde{\mathsf{P}}$ are nonnegative matrices with spectral radius less than one and zeros on the diagonal. Given $(\mathsf{x}, \chi) \in C^J \times C^{J\times J}$ with χ componentwise nonnegative and nondecreasing, we say that $(\mathsf{z}, \mathsf{y}, \mathsf{a}, \mathsf{b}) \in C_+^J \times C_+^J \times \mathbb{D}_+^{J\times J} \times \mathbb{D}_+^{J\times J}$ satisfies the augmented Skorohod problem associated with $(\mathsf{R}, \tilde{\mathsf{R}})$ for (x, χ) if the following conditions are satisfied:*

1. $\mathsf{z}(t) = \mathsf{x}(t) + \mathsf{R}\mathsf{y}(t)$ *for* $t \geq 0$,

2. $\mathsf{y}(0) = 0$, y *is componentwise nondecreasing and*

$$\int_0^\infty \mathsf{z}(t) d\mathsf{y}(t) = 0.$$

3. $\mathsf{a}(t) = \chi(t) - \tilde{\mathsf{R}}\mathsf{b}(t)$ *for* $t \geq 0$,

4. $\mathsf{b}(0) = 0$, $\mathsf{b}(t) \geq 0$, b *is componentwise nondecreasing and, for* $j = 1, \ldots, J$,

$$\int_0^\infty \mathsf{z}(t) d\mathsf{b}^j(t) = 0. \tag{5.4.1}$$

5. *For* $i = 1, \ldots, J$ *and* $t \geq 0$, $\mathsf{z}_i(t) = 0$ *implies* $\mathsf{a}_i(t) = 0$.

Building on results from Mandelbaum and Ramanan [74], we show in Appendix B.2 that the augmented Skorohod problem has a unique solution. To interpret solutions to the augmented Skorohod problem, we found it easiest to think of the dynamics of $(\mathsf{z}, \mathsf{a}^j)$ for each $j = 1, \ldots, J$ separately. When z hits the face $\mathsf{z}_I = 0$, then a^j jumps to the face $\mathsf{a}_I^j = 0$ in the direction of the unique vector in the column space of $\tilde{\mathsf{R}}^I$ which brings it to that face. We refer to Figure 2 for an illustrative example in the two-dimensional case.

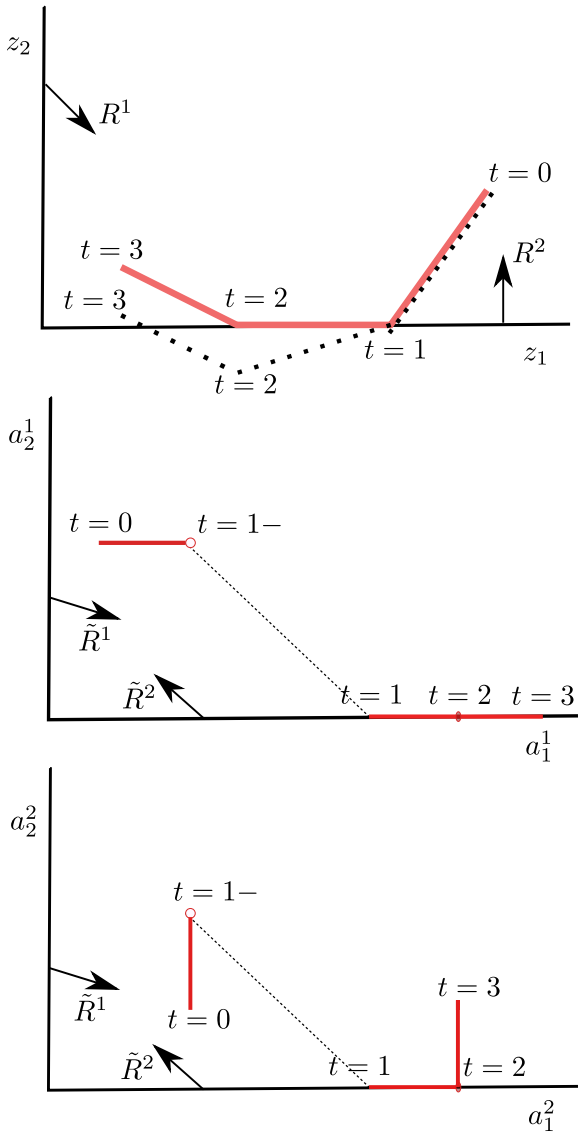

Figure 2: The first diagram depicts a trajectory of z, with corresponding 'free' path x (dotted). In the second and third diagram, the trajectories of a^1 and a^2 travel at unit rate right and up, respectively, until z hits $\partial\mathbb{R}^2_+$. The face $\mathsf{z}_2 = 0$ is hit at time $t = 1$, causing a^1 and a^2 to jump to the faces $\mathsf{a}^1_2 = 0$ and $\mathsf{a}^2_2 = 0$, respectively, in direction $\tilde{\mathsf{R}}^2$. Note that both $\mathsf{z}(0)$ and $\mathsf{a}(0) = \chi(0)$ are nonzero in these diagrams.

Unlike requirements 2 and 4 in Definition 5.3, requirement 5 is not a 'complementarity' condition. In view of the sample path dynamics in Figure 2, it may seem reasonable to replace requirement 5 by $\int_0^\infty \mathsf{a}^j(t)d\mathsf{y}(t) = 0$ or another complementarity condition between (y, z) and (a, b). In that case, however, the augmented Skorohod will fail to have a unique solution. This can be seen by verifying that both the left derivative and the right derivative of the reflection map satisfy $\int_0^\infty \mathsf{a}^j(t)d\mathsf{y}(t) = 0$ but only the left derivative (as defined in Definition 5.3) satisfies requirement 5.

We now make a connection between derivatives (sensitivity analysis) and solutions to the augmented Skorohod problem. Note that, unlike in Figure 2, one always has $\mathsf{a}(0) = \chi(0) = 0$ in this case.

Theorem 5.2 *Fix some* $\mathsf{x} \in C^J$*, and let* $\mathsf{z} = \Gamma(\mathsf{x})$ *and* $\mathsf{y} = \Phi(\mathsf{x})$ *be given by the oblique reflection map. Define the derivatives* $\mathsf{a} = \Gamma'(\mathsf{x})$ *and* $\mathsf{b} = -\Phi'(\mathsf{x})$ *as in Definition 5.2. Set* $\chi(t) = t\mathsf{l}$ *for* $t \geq 0$*. Then* $(\mathsf{z}, \mathsf{y}, \mathsf{a}, \mathsf{b})$ *satisfies the augmented Skorohod problem associated with* (R, R) *for* (x, χ)*.*

5.4.2 Stationary distribution of constrained diffusions and their derivatives

Our second main result specializes to diffusion processes and studies the stationary distribution of solutions to the augmented Skorohod problem. We show that it satisfies a generalized version of the basic adjoint relationship (BAR) for reflected Brownian motion. The proof relies on Ito's formula in conjunction with properties developed in the previous section. All results are formulated in terms of solutions to the augmented Skorohod problem, and the special case $\tilde{\mathsf{R}} = \mathsf{R}$ is of primary interest for the derivative process.

We first discuss the construction of constrained diffusion processes. We work with a d-dimensional standard Brownian motion $W = \{W(t) : t \geq 0\}$ adapted to some filtration $\{\mathcal{F}_t\}$, on an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We are given functions θ and σ on $\mathbb{R}^J_+$ taking values in $\mathbb{R}^J$ and $\mathbb{M}^{J\times d}$, respectively, which satisfy

the following standard Lipschitz and growth conditions: (1) For some $L < \infty$, we have $\|\sigma(\mathsf{x}) - \sigma(\mathsf{y})\|_2 + \|\theta(\mathsf{x}) - \theta(\mathsf{y})\|_2 \leq L\|\mathsf{x} - \mathsf{y}\|_2$ for all $\mathsf{x}, \mathsf{y} \in \mathbb{R}^J_+$. (2) For some $K < \infty$, we have $\|\theta(\mathsf{x})\|_2^2 + \|\sigma(\mathsf{x})\|_2^2 \leq K(1 + \|\mathsf{x}\|_2^2)$ for $\mathsf{x} \in \mathbb{R}^J_+$. Given any initial condition $\mathsf{Z}(0)$ with $\mathbb{E}\|\mathsf{Z}(0)\|_2^2 < \infty$, there exists a pathwise unique, strong solution $\{\mathsf{Z}(t) : t \geq 0\}$ to the stochastic differential equation with reflection (SDER)

$$d\mathsf{Z}(t) = \theta(\mathsf{Z}(t))dt + \sigma(\mathsf{Z}(t))dW(t) + \mathsf{R}d\mathsf{Y}(t). \tag{5.4.2}$$

This equation is shorthand for the statement that, almost surely, $\mathsf{Z} = \Gamma(\mathsf{X})$ and $\mathsf{X}(t) = \mathsf{Z}(0) + \int_0^t \theta(\mathsf{Z}(s))ds + \int_0^t \sigma(\mathsf{Z}(s))dW(s)$ for $t \geq 0$. Moreover, $\mathbb{E}\|\mathsf{Z}(t)\|_2^2$ is locally bounded as a function of t. For these and related results, see Anderson and Orey [6], Dupuis and Ishii [35], Karatzas and Shreve [59], Ramanan [88]. In particular, we have $\mathsf{Z}(t) \in \mathbb{R}^J_+$ for all $t \geq 0$. We define the diffusion matrix Σ through $\Sigma(\mathsf{z}) = \sigma(\mathsf{z})\sigma(\mathsf{z})'$ for $\mathsf{z} \in \mathbb{R}^J_+$. The special case of reflected Brownian motion follows upon taking constant functions σ and θ. Throughout this chapter, we only work with constrained diffusion processes that can be obtained through the oblique reflection map of Definition 5.1, and for which the time Z spends on $\partial\mathbb{R}^J_+$ has Lebesgue measure zero almost surely (this is only used in Section 5.7). Although the notions of SDER and their solutions can be defined more generally, our results cannot be extended to other settings using the present framework.

We next introduce an $\mathbb{M}_+^{J\times J}$-valued process $\mathsf{A} = \{\mathsf{A}(t) : t \geq 0\}$ through an augmented Skorohod problem. Although the special choice $\tilde{\mathsf{R}} = \mathsf{R}$ is most relevant for us given the connection with the derivative process, our treatment is not restricted to that case. Given some $\mathsf{A}(0)$, suppose that $(\mathsf{Z}, \mathsf{Y}, \mathsf{A}, \mathsf{B})$ satisfies the augmented Skorohod problem associated with $(\mathsf{R}, \tilde{\mathsf{R}})$ for (X, χ) with $\chi(t) = \mathsf{A}(0) + \mathsf{I}t$ and X as before. Also suppose $(\mathsf{Z}(0), \mathsf{A}(0))$ has some distribution u satisfying $\int \|\mathsf{z}\|_2^2 u(dz, da) < \infty$. This assumption guarantees existence of Z on a sample-path level, and therefore we do not need moment assumptions on $\mathsf{A}(0)$ in order to guarantee existence of the process A. The derivative process always starts at the origin (i.e., the zero matrix), but

here we have defined A with an arbitrary initial distribution since we are interested in stationary distributions for (Z, A). Recall that π is said to be a stationary distribution for (Z, A) if all marginal distributions of (Z, A) are π when $(\mathsf{Z}(0), \mathsf{A}(0))$ has distribution π, i.e., for every bounded measurable function $f : \mathbb{R}_+^J \times \mathbb{M}^{J\times J} \to \mathbb{R}$ and for every $t \geq 0$,

$$\mathbb{E}[f(\mathsf{Z}(t), \mathsf{A}(t))] = \int f(\mathsf{z}, \mathsf{a})\pi(d\mathsf{z}, d\mathsf{a}). \tag{5.4.3}$$

In view of Theorem 5.2, although a justification is outside the scope of this work, we think of the stationary distribution of (Z, A) with $\tilde{\mathsf{R}} = \mathsf{R}$ as the limiting distribution of Z jointly with its derivative process.

We remark that we still use the same notation π for stationary distributions as a convention. It should be clear that π refers to the stationary distribution for (Z, A) in this chapter, and it refers to the stationary distribution for $(\tilde{X}, \tilde{Z})$ in Part I of the thesis.

We define the following operators: Q_I is a projection operator with the following property. The matrix $\mathsf{Q}_I(\mathsf{a})$ is obtained from a by subtracting columns of $\tilde{\mathsf{R}}^I$, in such a way that the rows of $\mathsf{Q}_I(\mathsf{a})$ with indices in I become zero. That is, we have

$$\mathsf{Q}_I(\mathsf{a}) = \mathsf{a} - \tilde{\mathsf{R}}^I(\tilde{\mathsf{R}}_I^I)^{-1}\mathsf{a}_I, \tag{5.4.4}$$

where $\tilde{\mathsf{R}}_I^I$ is the principal submatrix of $\tilde{\mathsf{R}}$ obtained by removing rows and columns from $\tilde{\mathsf{R}}$ which do not lie in I. When $I = \emptyset$, we set $\mathsf{Q}_I(\mathsf{a}) = \mathsf{a}$ for $\mathsf{a} \in \mathbb{M}^{J\times J}$.

We also define operators L and T on $C_b^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$ through:

$$\begin{aligned} Lf(\cdot) &= \frac{1}{2}\langle \Sigma(\cdot), H_z f(\cdot)\rangle_{\text{HS}} + \langle \theta(\cdot), \nabla_{\mathsf{z}} f(\cdot)\rangle, \\ \mathsf{T}f(\cdot) &= Lf(\cdot) + \text{tr}(\nabla_{\mathsf{a}} f(\cdot)), \end{aligned} \tag{5.4.5}$$

where $\nabla_{\mathsf{z}} f$ and $H_z f$ denote the gradient and Hessian, respectively, with respect to the first argument of f, and we use $\nabla_{\mathsf{a}} f$ as discussed in Section 5.1. Thus, $\text{tr}(\nabla_{\mathsf{a}})$ is shorthand for $\sum_{i=1}^J d/d\mathsf{a}_{ii}$.

We can now formulate the following theorem, which is our second main result. We write I^c for the complement of a set I. We write z_I for the subvector of z consisting of the components with indices in I as before, and we also let $\mathsf{z}|_I$ denote the projection of z to $\{\mathsf{z} : \mathsf{z}_{I^c} = 0\}$.

Theorem 5.3 (Basic Adjoint Relationship) *Let the processes* Z *and* A *be defined as above, and suppose that* (Z, A) *has a unique stationary distribution* π *with* $\int(\|\mathsf{z}\|_2^2 + \|\mathsf{a}\|_1)\pi(dz, da) < \infty$. *Then there exists a finite Borel measure* ν *on* $\bigcup_i(F_i \cap F_i^{\mathsf{a}})$ *and, for* $I \subseteq \{1, \ldots, J\}$, *finite Borel measures* u_I *on* $(0,\infty)^{|I^c|} \times \mathbb{M}_+^{J\times J}$ *such that for any* $f \in C_{\mathsf{b}}^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$, *the following relationship holds:*

$$\int_{\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J}} \mathsf{T} f(\mathsf{z}, \mathsf{a}) d\pi(\mathsf{z}, \mathsf{a}) + \int_{\bigcup_i (F_i \cap F_i^{\mathsf{a}})} [\mathsf{R}' \nabla_{\mathsf{z}} f(\mathsf{z}, \mathsf{a})] d\nu(\mathsf{z}, \mathsf{a})$$
$$+ \sum_{I \subseteq \{1,\ldots,J\}: I \neq \emptyset} \int_{(0,\infty)^{|I^c|} \times \mathbb{M}_+^{J\times J}} [f(\mathsf{z}|_{I^c}, \mathsf{Q}_I(\mathsf{a})) - f(\mathsf{z}|_{I^c}, \mathsf{a})] du_I(\mathsf{z}_{I^c}, \mathsf{a}) = 0 \quad (5.4.6)$$

where the operators Q_I *and* T *are given in (5.4.4) and (5.4.5), respectively.*

Section 5.6.2 shows that the measures ν and u_I, $I \subseteq \{1, \ldots, J\}$ are completely determined by π, and expresses these measures in terms of π. We believe that (5.4.6) fully determines π, ν, and the u_I measures, but it is outside the scope of this chapter to prove this. For recent developments along these lines, see Dai and Dieker [23], Kang and Ramanan [58].

Theorem 5.3 does not have the same form as Theorem 5.1, and our next result brings these two forms closer. It is obtained by substituting a special class of functions in (5.4.6) so that the last term in (5.4.6) vanishes. To formulate the result, we need the following family of operators: for any $f \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}^{J\times J})$ and each set $I \subseteq \{1, 2, \ldots, J\}$, let

$$(O_I f)(\mathsf{z}, a) = \sum_{S \subseteq \{1,\ldots,J\} \setminus I} (-1)^{|S|} f(\Pi_{S\cup I}\mathsf{z}, \mathsf{Q}_I(\mathsf{a})), \quad (5.4.7)$$

$$O = \sum_{I \subseteq \{1,\ldots,J\}} O_I, \quad (5.4.8)$$

where $\Pi_{S\cup I}$ is the projection operator which sets the coordinates in $S \cup I$ equal to 0.

Corollary 5.1 *Let the processes* Z *and* A *be defined as above, and suppose that* (Z, A) *has a unique stationary distribution* π *with* $\int(\|\mathsf{z}\|_2^2 + \|\mathsf{a}\|_1)\pi(d\mathsf{z}, d\mathsf{a}) < \infty$. *Then there exists a finite Borel measure* ν *such that for any* $f \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$, *the following relationship holds:*

$$\int_{\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J}} [\mathsf{T} \circ Of](\mathsf{z}, \mathsf{a}) d\pi(\mathsf{z}, \mathsf{a}) + \int_{\bigcup_i (F_i \cap F_i^a)} [\mathsf{R}' \nabla_{\mathsf{z}} (Of)(\mathsf{z}, \mathsf{a})] d\nu(\mathsf{z}, \mathsf{a}) = 0, \quad (5.4.9)$$

where the operators T *and* O *are given in (5.4.5) and (5.4.8).*

We remark that the proof of this corollary shows that (5.4.9) is equivalent to several equations. That is, for any $f \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$ and each set $I \subseteq \{1, 2, \ldots, J\}$, π and ν must satisfy

$$\int_{\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J}} [\mathsf{T} \circ O_I f](\mathsf{z}, \mathsf{a}) d\pi(\mathsf{z}, \mathsf{a}) + \int_{\bigcup_i (F_i \cap F_i^a)} [\mathsf{R}' \nabla_{\mathsf{z}} (O_I f)(\mathsf{z}, \mathsf{a})] d\nu(\mathsf{z}, \mathsf{a}) = 0, (5.4.10)$$

where the operators O_I are defined in (5.4.7). Note that (5.4.10) produces 2^J equations, one of which is trivial. We refer to (5.4.10) as BAR_I.

We first check that (5.4.9) yields the classical BAR for the stationary distribution of the reflected Brownian motion Z when choosing $f(\mathsf{z}, \mathsf{a}) \equiv g(\mathsf{z})$ for some smooth g. One readily checks that in this case,

$$(Of)(\mathsf{z}, a) = \sum_{I \subseteq \{1,2,\ldots,J\}} \sum_{S \subseteq \{1,\ldots,J\} \setminus I} (-1)^{|S|} g(\Pi_{S\cup I} \mathsf{z}) = g(\mathsf{z}).$$

Substituting the above equation in (5.4.9), we immediately obtain the well-known basic adjoint relationship as introduced in Harrison and Williams [52] for reflected Brownian motion:

$$\int_{\mathbb{R}_+^J} Lg(\mathsf{z}) d\pi(\mathsf{z}) + \int_{\bigcup_i F_i} [\mathsf{R}' \nabla_{\mathsf{z}} g(\mathsf{z})] d\nu(\mathsf{z}) = 0, \quad (5.4.11)$$

where $d\pi(\mathsf{z}) = \int_{\mathsf{a} \in \mathbb{M}^{J\times J}} d\pi(\mathsf{z}, \mathsf{a})$ is the stationary distribution for Z and the Borel measure $d\nu(\mathsf{z})$ is given by $d\nu(\mathsf{z}) = \int_{\mathsf{a} \in \mathbb{M}^{J\times J}} d\nu(\mathsf{z}, \mathsf{a})$.

We next specialize (5.4.9) to the one-dimensional case, and we verify that we recover Theorem 5.1. This shows in particular that (5.4.9) fully determines π if $J = 1$. Indeed, it is readily seen that

$$(Of)(\mathsf{z}, \mathsf{a}) = (O_\emptyset f)(\mathsf{z}, \mathsf{a}) + (O_{\{1\}} f)(\mathsf{z}, \mathsf{a}) = f(\mathsf{z}, \mathsf{a}) - f(0, \mathsf{a}) + f(0, 0).$$

Combining this with (5.4.9) gives (5.2.4), but with $-\partial/\partial \mathsf{z} f(0,0)\theta$ replaced with $c\partial/\partial \mathsf{z} f(0,0)$ for some constant $c = \nu(\{0,0\}) > 0$. One can further show that $c = -\theta$, but we suppress the argument.

We next argue that none of the $2^J - 1$ nontrivial equations in (5.4.10) can be dropped, but we leave open the question whether they characterize π. We do so by illustrating the interplay between the different BAR_I in a simple example. Let $J = 3$ and consider $\mathsf{Z} = (\mathsf{Z}_1, \mathsf{Z}_2, \mathsf{Z}_3)$, where Z_1, Z_2, and Z_3 are three independent one-dimensional standard reflected Brownian motions. We do not need the second argument A, and therefore we make no distinction between (5.4.10) and a 'classical' analog of BAR_I in (5.4.10). This classical analog is obtained by considering (5.4.10) for f that do not depend on the second argument a, cf. how (5.4.11) was obtained from (5.4.9). The process Z has a unique stationary distribution π, which is a product form (see, e.g., Harrison and Williams [51] for details). $\mathrm{BAR}_{\{1,2\}}$ is equivalent with the third marginal distribution of π being exponential, with similar conclusions for $\mathrm{BAR}_{\{1,3\}}$ and $\mathrm{BAR}_{\{2,3\}}$. On the other hand, $\mathrm{BAR}_\emptyset$ and $\mathrm{BAR}_{\{j\}}$ for any $j \in \{1,2,3\}$ contain no information on the marginal distributions, in the sense that $O_\emptyset g = 0$ and $O_{\{j\}} g = 0$ for functions of the form $g(\mathsf{z}) = f_1(\mathsf{z}_1) + f_2(\mathsf{z}_2) + f_3(\mathsf{z}_3)$ (assuming appropriate smoothness). Still, $\mathrm{BAR}_{\{1\}}$ with $\mathrm{BAR}_{\{1,2\}}$ and $\mathrm{BAR}_{\{1,3\}}$ together imply that the push-forward of π under the projection map onto the last two coordinates has a product form solution since the two-dimensional reflected Brownian motion $(\mathsf{Z}_2, \mathsf{Z}_3)$ satisfies the so-called skew-symmetry condition, see Harrison and Williams [51, Theorem. 6.1] and Williams [108, Theorem. 1.2]. Consequently, one can think of $\mathrm{BAR}_{\{1\}}$ as describing the dependencies between the second and third component of π, with

marginal distributions determined by $\mathrm{BAR}_{\{1,2\}}$ and $\mathrm{BAR}_{\{1,3\}}$, respectively. Similarly, $\mathrm{BAR}_{\emptyset}$ describes the dependencies of the three two-dimensional push-forward measures of π.

5.5 Characteristics of derivatives and proof of Theorem 5.2

In this section, we prove Theorem 5.2. We also collect additional sample path properties of derivatives, with an emphasis on their jump behavior. These properties will be used in the proof of Theorem 5.3.

Throughout this section, we work under the conditions of Theorem 5.2. That is, we assume that $\mathsf{x} \in C^J$ is given and we write $\mathsf{z} = \Gamma(\mathsf{x})$, $\mathsf{y} = \Phi(\mathsf{x})$, $\mathsf{a} = \Gamma'(\mathsf{x})$ and $\mathsf{b} = -\Phi'(\mathsf{x})$. We also set $\chi(t) = t\mathsf{I}$ for $t \geq 0$.

5.5.1 Complementarity

This section connects the augmented Skorohod problem associated with (R, R) for (x, χ) with (z, a). Note that, in view of Definitions 5.1 and 5.2, the first two requirements of the augmented Skorohod problem in Definition 5.3 are immediately satisfied for $(\mathsf{x}, \mathsf{y}, \mathsf{z})$. It is immediate that $\mathsf{a} = \chi - \mathsf{R}\mathsf{b}$ by definition of a, so we must indeed choose $\tilde{\mathsf{R}} = \mathsf{R}$. We proceed with showing that a and b lie in $\mathbb{D}_+^{J\times J}$ as required for the augmented Skorohod problem, but it is convenient to first establish part of the fourth requirement.

Lemma 5.1 *The $\mathbb{M}^{J\times J}$-valued function b is componentwise nonnegative and nondecreasing.*

Proof Since $\chi(t) = t\mathsf{I}$ for $t \geq 0$, χ is evidently nonnegative and nondecreasing. The monotonicity result in Theorem 6 of Kella and Whitt [61] shows that for any fixed $\epsilon > 0$, each component of b_ϵ is nonnegative and nondecreasing. The lemma follows from the fact that b is the pointwise limit of the sequences $\{\mathsf{b}_\epsilon\}$ as $\epsilon \to 0+$.

Lemma 5.2 *The $\mathbb{M}^{J\times J}$-valued functions a and b lie in $\mathbb{D}_+^{J\times J}$.*

Proof Since b is nonnegative in view of Lemma 5.1, we will have shown the claim for b if we verify that $\mathsf{b} \in \mathbb{D}^{J\times J}$. We deduce from Theorem 1.1 in Mandelbaum and Ramanan [74] that each component of b is upper semicontinuous and that it has left and right limits everywhere. Since b is nondecreasing by Lemma 5.1, these properties imply that $\mathsf{b} \in \mathbb{D}^{J\times J}$.

We next show that $\mathsf{a} \in \mathbb{D}_+^{J\times J}$. Clearly, since $\mathsf{b} \in \mathbb{D}_+^{J\times J}$, we only need to show that a is nonnegative. Again by the monotonicity result in Theorem 6 of Kella and Whitt [61], for any fixed $\epsilon > 0$, each component of a_ϵ is nonnegative. This completes the proof of the lemma after letting $\epsilon \to 0+$.

We next investigate the fourth and fifth requirement of Definition 5.3. To this end, we need a characterization of b which relies heavily on [74].

Lemma 5.3 b *is the unique solution to the following system of equations: for* $i, j = 1, \ldots, J$, *and* $t \geq 0$,

$$\mathsf{b}_i^j(t) = \sup_{s\in\Phi_{(i)}(t)} [\delta_i^j s + [\mathsf{P}'\mathsf{b}^j]_i(s)],$$

where the supremum over an empty set should be interpreted as zero and

$$\Phi_{(i)}(t) \;=\; \{s \in [0,t] : \mathsf{z}_i(s) = 0\}. \tag{5.5.1}$$

Proof We use Theorem 1.1 of Mandelbaum and Ramanan s[74], which can be simplified in view of Lemma 5.1 and the nonnegativity of the matrix P. This theorem states that

$$\mathsf{b}_i^j(t) = \begin{cases} 0 & \text{if } t \in (0, t_{(i)}), \\ \sup_{s\in\Psi_{(i)}(t)}[\delta_i^j s + [\mathsf{P}'\mathsf{b}^j]_i(s)] & \text{if } t \in [t_{(i)}, \infty), \end{cases} \tag{5.5.2}$$

where $t_{(i)} = \inf\{t \geq 0 : \mathsf{z}_i(t) = 0\}$ and $\Psi_{(i)}(t) = \{s \in [0,t] : \mathsf{z}_i(s) = 0, \mathsf{y}_i(s) = \mathsf{y}_i(t)\}$. Observe that, again using Lemma 5.1, the supremum must be attained at the rightmost end of the closed interval $\Psi_{(i)}(t)$. Since y is nondecreasing and $\int_0^t \mathsf{z}_i(s)dy_i(s) = 0$, this is also the rightmost point of the closed set $\Phi_{(i)}(t)$. This establishes the lemma in view of the convention used for the supremum of an empty set.

Lemma 5.4 *Fix any $j = 1, \ldots, J$, we have*

$$\int_0^\infty \mathsf{z}(t) d\mathsf{b}^j(t) = 0. \tag{5.5.3}$$

Proof *Fix some $i = 1, \ldots, J$. Note that if $\mathsf{z}_i(t) > 0$ at time t, we deduce from the path continuity of z that there exists some $\epsilon > 0$ such that $\mathsf{z}_i(s) > 0$ for $s \in (t - \epsilon, t + \epsilon)$. This implies that $\Phi_{(i)}(s)$ is constant as a set-valued function for $s \in (t - \epsilon, t + \epsilon)$. Thus $\mathsf{b}_i(s)$ is constant for $s \in (t - \epsilon, t + \epsilon)$ by (5.5.2). Since i is arbitrary, this yields (5.5.3).*

Lemma 5.5 *If $\mathsf{z}_i(t) = 0$ for some i, then we have $\mathsf{a}_i(t) = 0$.*

Proof Suppose $\mathsf{z}_i(t) = 0$. In view of Lemma 5.1, we deduce from (5.5.2) that, for any $j = 1, \ldots, J$,

$$\mathsf{b}_i^j(t) = \delta_i^j t + [\mathsf{P}'\mathsf{b}^j]_i(t).$$

Now it follows from (5.3.2) and $\mathsf{R} = \mathsf{I} - \mathsf{P}'$ that

$$\mathsf{a}_i^j(t) = \delta_i^j t - [\mathsf{R}b^j]_i(t) = \delta_i^j t - \mathsf{b}_i^j(t) + [\mathsf{P}'\mathsf{b}^j]_i(t) = 0,$$

which completes the proof of the lemma.

The above two lemmas together with Lemma 5.2 yield two further complementarity conditions.

Corollary 5.2 *For any $j = 1, \ldots, J$, we have*

$$\begin{aligned} \int_0^\infty \mathsf{a}^j(t) d\mathsf{y}(t) &= 0, \\ \int_0^\infty \mathsf{a}^j(t) d\mathsf{b}^j(t) &= 0. \end{aligned} \tag{5.5.4}$$

Proof of Theorem 5.2 The claim is now immediate from (5.3.1) in conjunction with Lemmas 5.1, 5.2, 5.4, and 5.5.

5.5.2 Jumps of a

In this section, we collect sample path properties of a related to its jump behavior. This plays a critical role in the derivation of Theorem 5.3 and Corollary 5.1.

The next lemma states that a is linear whenever z is in the interior of $\mathbb{R}^J_+$.

Lemma 5.6 *If $\mathsf{z}(t) \in \mathbb{R}^J_+ \backslash \partial\mathbb{R}^J_+$ for $t \in [\alpha, \beta]$, then we have for $t \in [\alpha, \beta]$*

$$\mathsf{a}(t) = \mathsf{a}(\alpha) + (t - \alpha)\mathsf{l}.$$

In particular, a is continuous on (α, β) and can only have jumps when $\mathsf{z} \in \partial\mathbb{R}^J_+$.

Proof In view of (5.3.2), it suffices to show that b is constant for $t \in [\alpha, \beta]$. Since $\mathsf{z}(t) \in \mathbb{R}^J_+ \backslash \partial\mathbb{R}^J_+$ for $t \in [\alpha, \beta]$, we obtain from (5.5.1) that for each $i = 1, \ldots, J$, $\Phi_{(i)}(t)$ is constant as a set-valued function. Therefore, we deduce from (5.5.2) that $\mathsf{b}(t)$ is a constant in $\mathbb{M}^{J\times J}$ for $t \in [\alpha, \beta]$. The proof of the lemma is complete.

For any function g on $\mathbb{R}_+$, we write $\Delta g(t) = g(t) - g(t-)$. In view of the above lemma, we can characterize the continuous part of the function a. Formally, we write

$$\mathsf{a}(t) = \mathsf{a}^c(t) + \mathsf{a}^d(t),$$

where

$$\mathsf{a}^d(t) = \sum_{s\le t} \Delta\mathsf{a}(s).$$

We have the following corollary.

Corollary 5.3 $\mathsf{a}^c(t) = \mathsf{a}(0) + t\mathsf{l}$ *for any $t \ge 0$.*

We next characterize the jump direction of a when a jump occurs.

Lemma 5.7 *Fix a nonempty set $I \subseteq \{1, 2, \ldots, J\}$ and some $t > 0$. Suppose that $\mathsf{z}_k(t) = 0$ for $k \in I$ and $\mathsf{z}_i(t) > 0$ for $i \notin I$. If $\Delta\mathsf{a}(t) \neq 0$, then we must have*

$$\Delta\mathsf{a}(t) = -\sum_{k\in I} \mathsf{R}^k[\Delta\mathsf{b}]_k(t).$$

Proof Since $\mathsf{z}_i(t) > 0$ for $i \notin I$, we deduce from the sample path continuity of z that there exists some $\epsilon > 0$ such that for $i \notin I$, $\mathsf{z}_i(s) > 0$ for $s \in (t-\epsilon, t]$. This yields that for $i \notin I$, $\Phi_{(i)}(s)$ is a constant as a set-valued function for $s \in (t-\epsilon, t]$. From (5.5.2) we infer that for $i \notin I$, $\mathsf{b}_i(s)$ is constant for $s \in (t-\epsilon, t]$. This implies that $[\Delta \mathsf{b}]_i(t) = 0$ for $i \notin I$, and therefore that

$$\Delta \mathsf{a}(t) = -\mathsf{R}\Delta \mathsf{b}(t) = -\sum_{k=1}^{J} \mathsf{R}^k [\Delta \mathsf{b}]_k(t) = -\sum_{k \in I} \mathsf{R}^k [\Delta \mathsf{b}]_k(t).$$

This completes the proof of the lemma.

5.6 A basic adjoint relationship and proof of Theorem 5.3

This section is devoted to the proof of Theorem 5.3 and Corollary 5.1. The key idea is to apply Ito's formula to the semimartingale (Z, A) and use sample path properties of (Z, A) to analyze the stationary measure. This is a standard approach in the context of reflected Brownian motion, but the analysis here exposes new features due to the presence of jumps in the process A. Throughout, we work with the augmented filtration generated by W and $(\mathsf{Z}(0), \mathsf{A}(0))$.

5.6.1 Ito's formula for the semimartingale (Z, A)

In this section, we apply Ito's formula to the semimartingale (Z, A). We first show that (Z, A) is a semimartingale, i.e, each of its components is a semimartingale. Recall that a semimartingale is an adapted process which is the sum of a local martingale and a finite variation process, with sample paths in $\mathbb{D}$. For more details, we refer readers to Protter [86, Ch. 3] or Jacod and Shiryaev [56, Ch. 1].

Lemma 5.8 (Z, A) *is a semimartingale.*

Proof The process (Z, A) is adapted. This is a well-known property of Z, and $\mathsf{A}(t)$ is a deterministic functional of $\{\mathsf{Z}(s) : 0 \le s \le t\}$ and $\mathsf{A}(0)$ since it arises from an augmented Skorohod problem. We know from Lemma 5.2 that each component of

the process (Z, A) lies in $\mathbb{D}$. Since Z is a semimartingale, to show (Z, A) is a semimartingale, it suffices to show that A is a semimartingale. In fact, from Lemma 5.1 and (5.3.2) we immediately deduce that A is a finite variation process, that is, the paths of A are almost surely of finite variation on $[0, T]$ for any $T > 0$. In particular, A is a semimartingale.

By Ito's formula, e.g., Jacod and Shiryaev [56, Sec. I.4], we deduce from (5.4.2) that for any $f \in C^2_{\mathsf{b}}(\mathbb{R}^J_+ \times \mathbb{M}^{J\times J})$, we have

$$\begin{aligned} f(\mathsf{Z}(t), \mathsf{A}(t)) &= f(\mathsf{Z}(0), \mathsf{A}(0)) + \int_0^t [\sigma(\mathsf{Z}(s))' \nabla_{\mathsf{z}} f(\mathsf{Z}(s), \mathsf{A}(s-))] dW(s) \\ &\quad + \int_0^t [\mathsf{R}' \nabla_{\mathsf{z}} f(\mathsf{Z}(s), \mathsf{A}(s-))] dY(s) \\ &\quad + \int_0^t Lf(\mathsf{Z}(s), \mathsf{A}(s-)) ds + \int_0^t \nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s-)) dA^c(s) \\ &\quad + \sum_{s \le t} [f(\mathsf{Z}(s), \mathsf{A}(s)) - f(\mathsf{Z}(s), \mathsf{A}(s-))]. \end{aligned} \tag{5.6.1}$$

Compared to the formulation in Theorem I.4.57 of Jacod and Shiryaev [56], we have absorbed the last sum of the jump part into the integral $\int_0^t \nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s-)) dA^c(s)$. This is justified by noting that, since $\Delta \mathsf{A}(s) = -\tilde{\mathsf{R}} \Delta \mathsf{B}(s)$ for some nonnegative and (componentwise) nondecreasing process B according to Definition 5.3,

$$\sum_{s \le t} \|\Delta \mathsf{A}(s)\|_1 \le C \sum_{s \le t} \|\Delta \mathsf{B}(s)\|_1 = C \|\mathsf{B}(t)\|_1 < \infty, \tag{5.6.2}$$

where C denotes some constant depending on $\tilde{\mathsf{R}}$. Note that this also implies that the last term on the right-hand side of (5.6.1) is absolutely convergent. Indeed, combining the above bound with $f \in C^2_b(\mathbb{R}^J_+ \times \mathbb{M}^{J\times J})$ yields $\sum_{s \le t} |f(\mathsf{Z}(s), \mathsf{A}(s)) - f(\mathsf{Z}(s), \mathsf{A}(s-))| < \infty$.

Suppose that (Z, A) is positive recurrent and has a unique stationary distribution π. Henceforth we assume that $(\mathsf{Z}(0), \mathsf{A}(0))$ has distribution π, and we write $\mathbb{E}_\pi$ instead of $\mathbb{E}$. After taking an expectation with respect to π on both sides of (5.6.1), the term involving dW vanishes since it is a martingale term. We next analyze the second

to last term on the right-hand side. From Corollary 5.3 and the fact that A has countably many jumps (Lemma 5.1), we deduce that

$$\begin{aligned}\mathbb{E}_\pi \int_0^t \nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s-)) dA^c(s) &= \mathbb{E}_\pi \int_0^t \nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s-)) d(s\mathsf{I}) \\ &= \mathbb{E}_\pi \int_0^t \nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s)) d(s\mathsf{I}) \\ &= \mathbb{E}_\pi \int_0^t \mathrm{tr}(\nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s))) ds.\end{aligned}$$

Since $f \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}^{J\times J})$, we have from Fubini's theorem and the definition of stationarity in (5.4.3) that

$$\begin{aligned}\mathbb{E}_\pi \int_0^t \mathrm{tr}(\nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s))) ds &= \int_0^t \mathbb{E}_\pi \mathrm{tr}(\nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s))) ds \\ &= t \int \mathrm{tr}(\nabla_{\mathsf{a}} f(\mathsf{z}, \mathsf{a})) d\pi(\mathsf{z}, \mathsf{a}).\end{aligned}$$

Thus we obtain

$$\mathbb{E}_\pi \int_0^t \nabla_{\mathsf{a}} f(\mathsf{Z}(s), \mathsf{A}(s-)) dA^c(s) = t \int \mathrm{tr}(\nabla_{\mathsf{a}} f(\mathsf{z}, \mathsf{a})) d\pi(\mathsf{z}, \mathsf{a}).$$

A similar argument applies to the fourth term on the right-hand side of (5.6.1). We conclude that, for each $t \geq 0$ and each $f \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}^{J\times J})$,

$$\begin{aligned}0 &= t \int [Tf(\mathsf{z}, \mathsf{a})] d\pi(\mathsf{z}, \mathsf{a}) + \mathbb{E}_\pi \int_0^t [\mathsf{R}' \nabla_{\mathsf{z}} f(\mathsf{Z}(s), \mathsf{A}(s-))] dY(s) \\ &\quad + \mathbb{E}_\pi \sum_{s \leq t} [f(\mathsf{Z}(s), \mathsf{A}(s)) - f(\mathsf{Z}(s), \mathsf{A}(s-))], \end{aligned} \tag{5.6.3}$$

where T is given in (5.4.5). This equation serves as the starting point for proving Theorem 5.3.

5.6.2 The boundary term

In this section we rewrite the boundary term in (5.6.3), i.e., the term involving $d\mathsf{Y}$. Let $\nu = (\nu_1, \ldots, \nu_J)$ be the unique vector of measures on $\partial\mathbb{R}_+^J \times \mathbb{M}^{J\times J}$ for which

$$\int h(\mathsf{z}, \mathsf{a}) \nu(dz, da) = \mathbb{E}_\pi \int_0^1 h(\mathsf{Z}(s), \mathsf{A}(s-)) d\mathsf{Y}(s),$$

for all continuous $h : \partial\mathbb{R}^J_+ \times \mathbb{M}^{J\times J} \to \mathbb{R}^J$ with compact support. This is a well-defined measure by the following lemma. For a different proof in the reflected Brownian motion case, see Harrison and Williams [52, Section 8].

Lemma 5.9 *We have* $\mathbb{E}_\pi \mathsf{Y}(1) < \infty$ *componentwise.*

Proof Since $\mathsf{Y}(1) \geq 0$, it is enough to show that $\mathbb{E}_\pi \|R\mathsf{Y}(1)\|_1 < \infty$. We prove the stronger statement that $\mathbb{E}_\pi \|R\mathsf{Y}(1)\|_2^2 < \infty$. From the fact that Z satisfies the SDER (5.4.2), we obtain

$$\mathbb{E}_\pi \left\|\mathsf{RY}(1)\right\|_2^2 = \mathbb{E}_\pi \left\| \mathsf{Z}(1) - \mathsf{Z}(0) - \int_0^1 \theta(\mathsf{Z}(s))ds - \int_0^1 \sigma(\mathsf{Z}(s))dW(s) \right\|_2^2.$$

Since the mapping $t \mapsto \mathbb{E}_\pi \|\mathsf{Z}(t)\|_2^2$ is locally bounded, we deduce from the growth condition on θ that $\mathbb{E}_\pi \left\| \int_0^1 \theta(\mathsf{Z}(s))ds \right\|_2^2 < \infty$. Similarly, we have

$$\mathbb{E}_\pi \left\| \int_0^1 \sigma(\mathsf{Z}(s))dW(s) \right\|_2^2 = \mathbb{E}_\pi \int_0^1 \mathrm{tr}\Sigma(\mathsf{Z}(s))ds < \infty,$$

where the finiteness follows from the growth condition on σ.

Our next goal is to give a characterization of measure ν in terms of π, which we carry out through Laplace transforms. We start with determining the support of ν.

Lemma 5.10 *The support of* ν *is* $\bigcup_i (F_i \cap F_i^{\mathsf{a}})$.

Proof In view of Lemma 5.2, it is clear that A can have at most countably many jumps. For any continuous $h : \partial\mathbb{R}^J_+ \times \mathbb{M}^{J\times J} \to \mathbb{R}^J$ with compact support, we have

$$\int_0^1 h(\mathsf{Z}(s), \mathsf{A}(s-))d\mathsf{Y}(s) = \int_0^1 h(\mathsf{Z}(s), \mathsf{A}(s))d\mathsf{Y}(s),$$

since the measure $d\mathsf{Y}$ is continuous and the integrand has countably many jumps by Lemma 5.1. It follows from the definition of ν that

$$\int h(\mathsf{z}, \mathsf{a})\nu(dz, da) = \mathbb{E}_\pi \int_0^1 h(\mathsf{Z}(s), \mathsf{A}(s))d\mathsf{Y}(s). \tag{5.6.4}$$

The complementarity conditions $\int_0^\infty \mathsf{Z}(t)d\mathsf{Y}(t) = 0$ and (5.5.4) imply the lemma.

On combining Equations (5.6.3) and (5.6.4) we obtain that for any $f \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$,

$$\begin{aligned} 0 &= \int_{\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J}} [\mathsf{T}f(\mathsf{z},\mathsf{a})] d\pi(\mathsf{z},\mathsf{a}) + \int_{\bigcup_i (F_i \cap F_i^{\mathsf{a}})} [\mathsf{R}'\nabla_{\mathsf{z}} f(\mathsf{z},\mathsf{a})] d\nu(\mathsf{z},\mathsf{a}) \\ &\quad + \frac{1}{t}\mathbb{E}_\pi \sum_{s\le t} [f(\mathsf{Z}(s),\mathsf{A}(s)) - f(\mathsf{Z}(s),\mathsf{A}(s-))]. \end{aligned} \tag{5.6.5}$$

We now express the Laplace transform of ν in terms of the Laplace transform of π. Set $f(\mathsf{z},\mathsf{a}) = \exp(-\eta'\mathsf{z} - \langle \alpha, \mathsf{a}\rangle_{\text{HS}}) \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$ where $(\eta,\alpha) \in \mathbb{R}_+^J \times \mathbb{M}_+^{J\times J}$. After substituting f in (5.6.5), we obtain

$$\pi^*(\eta,\alpha) - \sum_{j=1}^J (\mathsf{R}'\eta)_j \nu_j^*(\eta,\alpha) + H(\eta,\alpha) = 0, \tag{5.6.6}$$

where

$$\begin{aligned} \pi^*(\eta,\alpha) &= \int_{\mathbb{R}_+^J \times \mathbb{M}^{J\times J}} \left[\frac{1}{2}\eta'\Sigma(\mathsf{z})\eta + \eta'\theta(\mathsf{z}) - \sum_{i=1}^J \alpha_i\right] e^{-\eta\cdot\mathsf{z} - \alpha\cdot\mathsf{a}} d\pi(\mathsf{z},\mathsf{a}), \\ \nu_j^*(\eta,\alpha) &= \int_{F_j \cap F_j^{\mathsf{a}}} e^{-\eta\cdot\mathsf{z}-\alpha\cdot\mathsf{a}} d\nu_j(\mathsf{z},\mathsf{a}), \\ H(\eta,\alpha) &= \mathbb{E}_\pi \sum_{s\le 1} [e^{-\eta\cdot\mathsf{Z}(s)} \cdot (e^{-\alpha\cdot\mathsf{A}(s)} - e^{-\alpha\cdot\mathsf{A}(s-)})]. \end{aligned}$$

Dividing (5.6.6) by $\eta_j > 0$ and letting $\eta_j \to \infty$, we deduce that

$$\nu_j^*(\eta,\alpha) = \frac{1}{2}\lim_{\eta_j\to\infty} \eta_j \int_{\mathbb{R}_+^J \times \mathbb{M}^{J\times J}} \Sigma_{jj}(\mathsf{z}) e^{-\eta\cdot\mathsf{z}-\alpha\cdot\mathsf{a}} d\pi(\mathsf{z},\mathsf{a}), \tag{5.6.7}$$

where we have used the fact that $\nu_j(F_j \cap F_i) = 0$ for $i \neq j$ so that $\lim_{\eta_j\to\infty} \nu_i^*(\eta,\alpha) = 0$ by the dominated convergence theorem. Since all terms in (5.6.6) vanish in the limit by dominated convergence except for the term with ν_j^* and the term with π^*, existence of the limit in (5.6.7) follows immediately from the fact that $\nu_j(\eta,\alpha)$ does not depend on η_j. Under further regularity conditions on π, one can use the initial value theorem for Laplace transforms to show that $d\nu_j = \frac{1}{2}\Sigma_{jj} d\pi_j$ for an appropriate restriction π_j of π. Carrying out this procedure provides little additional insight, and we therefore suppress further details.

5.6.3 The jump term

We now proceed to investigate the jump term, i.e., the term in (5.6.3) involving the countable sum. Lemma 5.6 implies that jumps in A can only occur when Z lies hits the boundary $\partial\mathbb{R}^J_+$ of the nonnegative orthant, which motivates the following definition. For $I \subseteq \{1,\dots,J\}, I \neq \emptyset$, we define measures u_I on $\mathbb{R}_+^{|I^c|} \times \mathbb{M}_+^{J\times J}$ with support in $(0,\infty)^{|I^c|} \times \mathbb{M}_+^{J\times J}$. We set, for Borel sets $G \subseteq (0,\infty)^{|I^c|}, C \subseteq \mathbb{M}_+^{J\times J}$,

$$u_I(G,C) = \mathbb{E}_\pi \sum_{s\leq 1:\mathsf{Z}_I(s)=0,\mathsf{Z}_{I^c}(s)\in G,\mathsf{A}(s)\neq\mathsf{A}(s-)} 1_C\{\mathsf{A}(s-)\}.$$

This is a well-defined σ-finite measure because of (5.6.2) and $\mathbb{E}_\pi\|\mathsf{B}(1)\|_1 = \mathbb{E}_\pi\|\mathsf{A}(1)-\mathsf{A}(0)-E\|_1 \leq 2\mathbb{E}_\pi\|\mathsf{A}(0)\|_1+J < \infty$, so that $\mathbb{E}_\pi|\sum_{s\leq 1}[f(\mathsf{Z}(s),\mathsf{A}(s))-f(\mathsf{Z}(s),\mathsf{A}(s-))]| < \infty$ for $f \in C_b^2(\mathbb{R}^J_+ \times \mathbb{M}^{J\times J})$. It is possible to express these measures in terms of π using the theory of distributions; this is done in Section 5.7.

The primary objective of this subsection is to show that the jump term in (5.6.3) vanishes for a special class of functions, which is key in our proof of Corollary 5.1. Throughout, we fix a set $I \subseteq \{1,\dots,J\}$. Recall the definition of O_I in (5.4.7). It is our aim to show that the jump term vanishes for functions of the form $O_I f$, where $f \in C_b^2(\mathbb{R}^J_+ \times \mathbb{M}^{J\times J})$ as before. We first introduce a lemma.

Lemma 5.11 *For any $f : \mathbb{R}^J_+ \times \mathbb{M}^{J\times J} \to \mathbb{R}$, if $\mathsf{z}_j = 0$ for some $j \notin I$, then for any $\mathsf{a} \in \mathbb{M}^{J\times J}$ we have*

$$\sum_{S\subseteq\{1,\dots,J\}\setminus I} (-1)^{|S|} f(\Pi_{S\cup I}\mathsf{z},\mathsf{a}) = 0.$$

In particular, if $\mathsf{z}_j = 0$ for some $j \notin I$, then we have $O_I f(\mathsf{z},\mathsf{a}) = 0$.

Proof Suppose $\mathsf{z}_j = 0$ for some $j \notin I$. Then for any set $S\subseteq \{1,\dots,J\}\setminus I$ with $j \notin S$,

we have $\Pi_{S\cup I}\mathsf{z} = \Pi_{S\cup I\cup\{j\}}\mathsf{z}$. Using this observation, we deduce that

$$\begin{aligned}
&\sum_{S\subseteq\{1,\ldots,J\}\setminus I}(-1)^{|S|}f(\Pi_{S\cup I}\mathsf{z},\mathsf{a})\\
&= \sum_{S\subseteq\{1,\ldots,J\}\setminus I:j\in S}(-1)^{|S|}f(\Pi_{S\cup I}\mathsf{z},\mathsf{a}) + \sum_{S\subseteq\{1,\ldots,J\}\setminus I:j\notin S}(-1)^{|S|}f(\Pi_{S\cup I}\mathsf{z},\mathsf{a})\\
&= \sum_{S\subseteq\{1,\ldots,J\}\setminus I:j\in S}(-1)^{|S|}f(\Pi_{S\cup I}\mathsf{z},\mathsf{a}) + \sum_{S\subseteq\{1,\ldots,J\}\setminus I:j\notin S}(-1)^{|S|}f(\Pi_{S\cup I\cup\{j\}}\mathsf{z},\mathsf{a})\\
&= \sum_{S\subseteq\{1,\ldots,J\}\setminus I:j\in S}(-1)^{|S|}f(\Pi_{S\cup I}\mathsf{z},\mathsf{a}) + \sum_{\tilde{S}\subseteq\{1,\ldots,J\}\setminus I:j\in\tilde{S}}(-1)^{|\tilde{S}|-1}f(\Pi_{\tilde{S}\cup I}\mathsf{z},\mathsf{a})\\
&= 0.
\end{aligned}$$

The proof of the lemma is complete.

Now we are ready to show that the jump term vanishes for functions of the form $O_I f$. For any $K\subseteq\{1,\ldots,J\}$, Z_K denotes the process whose components are those of Z with indices in K.

Lemma 5.12 *For each $t\geq 0$ and any measurable $f:\mathbb{R}_+^J\times\mathbb{M}^{J\times J}\to\mathbb{R}$, we have*

$$\mathbb{E}_\pi\sum_{s\leq t}[O_I f(\mathsf{Z}(s),\mathsf{A}(s)) - O_I f(\mathsf{Z}(s),\mathsf{A}(s-))] = 0. \tag{5.6.8}$$

Proof By Lemma 5.6 and Lemma 5.11, we have

$$\begin{aligned}
&\mathbb{E}_\pi\sum_{s\leq t}[O_I f(\mathsf{Z}(s),\mathsf{A}(s)) - O_I f(\mathsf{Z}(s),\mathsf{A}(s-))]\\
&= \sum_{\emptyset\neq K\subseteq\{1,\ldots,J\}}\mathbb{E}_\pi\sum_{s\leq t:\mathsf{Z}_K(s)=0,\mathsf{Z}_{\{1,\ldots,J\}\setminus K}(s)>0}[O_I f(\mathsf{Z}(s),\mathsf{A}(s)) - O_I f(\mathsf{Z}(s),\mathsf{A}(s-))]\\
&= \sum_{\emptyset\neq K\subseteq I}\mathbb{E}_\pi\sum_{s\leq t:\mathsf{Z}_K(s)=0,\mathsf{Z}_{\{1,\ldots,J\}\setminus K}(s)>0}[O_I f(\mathsf{Z}(s),\mathsf{A}(s)) - O_I f(\mathsf{Z}(s),\mathsf{A}(s-))].
\end{aligned}$$

Therefore, to show (5.6.8) it suffices to show for each nonempty set $K\subseteq I$, we have

$$\mathbb{E}_\pi\sum_{s\leq t:\mathsf{Z}_K(s)=0,\mathsf{Z}_{\{1,\ldots,J\}\setminus K}(s)>0}[O_I f(\mathsf{Z}(s),\mathsf{A}(s)) - O_I f(\mathsf{Z}(s),\mathsf{A}(s-))] = 0. \tag{5.6.9}$$

To prove (5.6.9) we first deduce from Definition 5.3 that when $\mathsf{Z}_K(s)=0$ and $\mathsf{Z}_{\{1,\ldots,J\}\setminus K}(s)>0$,

$$\mathsf{Q}_K(\mathsf{A}(s-)) = \mathsf{A}(s).$$

Next, since $K \subseteq I$, we use the projection property of the operator Q_I to obtain

$$\mathsf{Q}_I(\mathsf{A}(s)) = \mathsf{Q}_I(\mathsf{Q}_K(\mathsf{A}(s-)) = \mathsf{Q}_I(\mathsf{A}(s-)).$$

Now (5.6.9) readily follows from the definition of O_I as in (5.4.7). Thus we have completed the proof of the lemma.

5.6.4 Proofs of Theorem 5.3 and Corollary 5.1

We now prove Theorem 5.3 and Corollary 5.1.

Proof of Theorem 5.3 We rewrite the jump term in (5.6.5) using the jump measures. In view of Lemmas 5.5 and 5.7,

$$\begin{aligned}
&\mathbb{E}_\pi \sum_{s\le 1}[f(\mathsf{Z}(s),\mathsf{A}(s)) - f(\mathsf{Z}(s),\mathsf{A}(s-))] \\
&= \sum_{\emptyset\neq K\subseteq\{1,\ldots,J\}} \mathbb{E}_\pi \sum_{s\le 1:\mathsf{Z}_K(s)=0,\mathsf{Z}_{K^c}(s)>0} [f(\mathsf{Z}|_{K^c}(s),\mathsf{A}(s)) - f(\mathsf{Z}|_{K^c}(s),\mathsf{A}(s-))] \\
&= \sum_{\emptyset\neq K\subseteq\{1,\ldots,J\}} \mathbb{E}_\pi \sum_{s\le 1:\mathsf{Z}_K(s)=0,\mathsf{Z}_{K^c}(s)>0,\mathsf{A}(s)\neq\mathsf{A}(s-)} [f(\mathsf{Z}|_{K^c}(s),\mathsf{Q}_K(\mathsf{A}(s-))) - f(\mathsf{Z}|_{K^c}(s),\mathsf{A}(s-))] \\
&= \sum_{\emptyset\neq K\subseteq\{1,\ldots,J\}} \int_{\mathsf{z}_{K^c},\mathsf{a}} [f(\mathsf{z}|_{K^c},\mathsf{Q}_K(\mathsf{a})) - f(\mathsf{z}|_{K^c},\mathsf{a})] du_K(\mathsf{z}_{K^c},\mathsf{a}).
\end{aligned}$$

Thus, Theorem 5.3 follows from (5.6.5).

Proof of Corollary 5.1 Equation (5.4.10) immediately follows from (5.6.5) and Lemma 5.12. Summing all the equations in (5.4.10) over the sets $I \subseteq \{1,\ldots,J\}$, we obtain (5.4.9).

5.7 Jump measures

In this section, we further investigate the jump term in (5.6.3), resulting in a characterization of jump measures u_I in terms of the stationary distribution π. We start with an auxiliary result on the measures u_I.

Lemma 5.13 *For each $I \subseteq \{1,\ldots,J\}, I \neq \emptyset$ and $k = 1,\ldots,J$, we have $u_I(\{(\mathsf{z}_{I^c},\mathsf{a}) : \mathsf{a}_k = 0\}) = 0$.*

Proof We exploit the dynamics of the augmented Skorohod problem. Since $\mathsf{A}_k(s-) = 0$ implies $\mathsf{Z}_k(s) = 0$, we have $u_I(\{(\mathsf{z}_{I^c}, \mathsf{a}) : \mathsf{a}_k = 0\}) = 0$ for $k \in I^c$. We next consider $k \in I$. Since the continuous part of A_k^k is strictly increasing when $\mathsf{Z}_k > 0$, the only possibility for $\mathsf{Z}_I(s) = 0$, $\mathsf{A}_k(s-) = 0$, and $\mathsf{A}(s) \neq \mathsf{A}(s-)$ to occur simultaneously is for Z to hit the face $\mathsf{z}_I = 0$ without having left the face $\mathsf{z}_k = 0$ for some positive amount of time. Since the time Z spends on the boundary has Lebesgue measure zero, this cannot happen almost surely.

To proceed with our description of the measures u_I, we need tools from theory of distributions (or generalized functions). For background on this theory, see Duistermaat and Kolk [34], Rudin [93]. For $I \subseteq \{1, \ldots, J\}$, we define the operator T_I^* on distributions through

$$\mathsf{T}_I^* f = \frac{1}{2} \sum_{i,j \in I} \frac{\partial^2}{\partial \mathsf{z}_i \partial \mathsf{z}_j} [\Sigma_{ij}(\cdot) f] - \sum_{j \in I} \theta_j \frac{\partial}{\partial \mathsf{z}_j} f - \mathrm{tr}(\nabla_{\mathsf{a}} f),$$

for any distribution f. With the understanding that we identify any probability measure with the distribution it generates, we can differentiate (probability) measures and T_I^* can act on measures. We also define

$$d\pi_I(\mathsf{z}_{I^c}, \mathsf{a}) = \int_{\mathsf{z}_I} d\pi(\mathsf{z}, \mathsf{a}).$$

The main result of this section is that u_I can be expressed in terms of π. Indeed, together with Lemma 5.13, it completely determines u_I.

Proposition 5.1 *For each $I \subseteq \{1, \ldots, J\}, I \neq \emptyset$, we have, with $\mathsf{z}_{I^c} \in (0, \infty)^{|I^c|}, \mathsf{a} \in \mathbb{M}_+^{J \times J}$ and $\mathsf{a}_k \neq 0$ for $k = 1, \ldots, J$,*

$$du_I(\mathsf{z}_{I^c}, \mathsf{a}) = \sum_{K \subseteq I, K \neq \emptyset} (-1)^{|I \setminus K|} \int_{\mathsf{z}_{I \setminus K}} [\mathsf{T}_{K^c}^* d\pi_K](\mathsf{z}_{K^c}, \mathsf{a}).$$

Proof Equation (5.6.5) forms the basis of the proof, together with the identity

$$\begin{aligned} &\mathbb{E}_\pi \sum_{s \leq 1} [f(\mathsf{Z}(s), \mathsf{A}(s)) - f(\mathsf{Z}(s), \mathsf{A}(s-))] \\ &\quad = \sum_{\emptyset \neq K \subseteq \{1, \ldots, J\}} \int_{\mathsf{z}_{K^c}, \mathsf{a}} [f(\mathsf{z}|_{K^c}, \mathsf{Q}_K(\mathsf{a})) - f(\mathsf{z}|_{K^c}, \mathsf{a})] du_K(\mathsf{z}_{K^c}, \mathsf{a}), \end{aligned}$$

which was established in Section 5.6.4. Fix some nonempty $I \subseteq \{1, \dots, J\}$. For $f \in C_b^2(\mathbb{R}_+^J \times \mathbb{M}_+^{J\times J})$ with the property that f vanishes on $\bigcup_{i\in I^c} F_i \cup \bigcup_i F_i^{\mathsf{a}}$, (5.6.5) reduces to

$$\int_{\mathbb{R}_+^J\times\mathbb{M}_+^{J\times J}} \mathsf{T}f(\mathsf{z},\mathsf{a})d\pi(\mathsf{z},\mathsf{a}) = \sum_{L\subseteq I: L\neq I} \int_{\mathsf{z}_{I^c\cup L},\mathsf{a}} f(\mathsf{z}|_{I^c\cup L},\mathsf{a}) du_{I\setminus L}(\mathsf{z}_{I^c\cup L},\mathsf{a}).$$

If moreover $f(\mathsf{z},\mathsf{a})$ does not depend on z_I, this can be simplified further:

$$\int_{\mathbb{R}_+^J\times\mathbb{M}_+^{J\times J}} \mathsf{T}f(\mathsf{z},\mathsf{a})d\pi(\mathsf{z},\mathsf{a}) = \sum_{L\subseteq I: L\neq I} \int_{\mathsf{z}_{I^c},\mathsf{a}} f(\mathsf{z}|_{I^c},\mathsf{a}) \int_{\mathsf{z}_L} du_{I\setminus L}(\mathsf{z}_{I^c\cup L},\mathsf{a}). \tag{5.7.1}$$

The left-hand side can be rewritten using the theory of differentiation for distributions Duistermaat and Kolk [34, Ch. 4] or Rudin [93, Sec. II.6.12]. This leads to

$$\int_{\mathbb{R}_+^J\times\mathbb{M}^{J\times J}} \mathsf{T}f(\mathsf{z},\mathsf{a})d\pi(\mathsf{z},\mathsf{a}) = \int_{\mathsf{z}_{I^c},\mathsf{a}} f(\mathsf{z}|_{I^c},\mathsf{a})[\mathsf{T}_{I^c}^* d\pi_I](\mathsf{z}_{I^c},\mathsf{a}).$$

Combining this with (5.7.1) and rearranging terms, we get

$$\begin{aligned}&\int_{\mathsf{z}_{I^c},\mathsf{a}} f(\mathsf{z}|_{I^c},\mathsf{a}) du_I(\mathsf{z}_{I^c},\mathsf{a})\\ &\quad= \int_{\mathsf{z}_{I^c},\mathsf{a}} f(\mathsf{z}|_{I^c},\mathsf{a})[\mathsf{T}_{I^c}^* d\pi_I](\mathsf{z}_{I^c},\mathsf{a}) - \sum_{L\subseteq I: L\neq\emptyset, L\neq I} \int_{\mathsf{z}_{I^c},\mathsf{a}} f(\mathsf{z}|_{I^c},\mathsf{a}) \int_{\mathsf{z}_L} du_{I\setminus L}(\mathsf{z}_{I^c\cup L},\mathsf{a}).\end{aligned}$$

This shows that, for $\mathsf{z}_{I^c} \in (0,\infty)^{|I^c|}$, $\mathsf{a} \in \mathbb{M}_+^{J\times J}$ and $\mathsf{a}_k \neq 0$ for $k = 1, \dots, J$,

$$du_I(\mathsf{z}_{I^c},\mathsf{a}) = \mathsf{T}_{I^c}^* d\pi_I(\mathsf{z}_{I^c},\mathsf{a}) - \sum_{L\subseteq I, L\neq\emptyset, L\neq I} \int_{\mathsf{z}_L} du_{I\setminus L}(\mathsf{z}_{I^c\cup L},\mathsf{a}).$$

Since $|I\setminus L| < |I|$, this representation allows us to finish the proof of the proposition by an elementary induction argument on $|I|$. Alternatively, one could use a version of the inclusion-exclusion principle, see, e.g., Stanley [99, Sec. 2.1].

STOCHASTIC MODELS FOR SERVICE SYSTEMS AND LIMIT ORDER BOOKS

PART III

Limit order books

by

Xuefeng Gao

CHAPTER VI

HYDRODYNAMIC LIMIT OF ORDER BOOK DYNAMICS

6.1 Introduction

In this chapter, we study limit order books. As a trading mechanism, limit order books have gained growing popularity in equity and derivative markets in the past two decades. Nowadays, the majority of the world's financial markets, such as Electronic Communication Networks in United States, the Hong Kong Stock Exchange and the Toronto Stock Exchange, are organized as electronic limit-order books to match buyers and sellers. There have also been intense research activities on limit order books, including both empirical and modelling studies. See, e.g., Parlour and Seppi [81], Gould et al. [47], Chakraborti et al. [18], Chakraborti et al. [19] for reviews and surveys.

Our work is motivated by a desire to better understand the interplay between order book shape evolution and optimal executions. The goal of this work is to characterize the transient behavior of order book shape dynamics on the "macroscopic" time scale (e.g., minutes). The shape of an order book describes the number of awaiting limit orders at each price level. Understanding its temporal evolution is critical for traders to optimally execute orders and reduce transaction costs, see, e.g., Obizhaeva and Wang [79], Alfonsi et al. [5], Predoiu et al. [85], Alfonsi and Acevedo [4] and Fruth et al. [38]. While the authors in those studies typically assume some exogenous and potentially time-varying shape density function in the absence of large trades, we are interested in an analytically tractable description of the order book shape using observable quantities such as order arrival and cancellation rates.

There have been quite a few microscopic dynamical models of limit order books, see, e.g., Bak et al. [8], Smith et al. [98], Cont et al. [21] and Abergel and Jedidi [1]. Our work is based on Cont et al. [21], where the authors proposed to use continuous-time Markov chain to capture the short-term dynamics of limit order books. In their model, there are finite number of security prices and the state of the order book is described by a vector whose element is the order quantity offered in the order book at each price. Market order arrivals, limit order placement and cancelations are governed by independent Poisson processes. Using Laplace transform analysis, the authors were able to efficiently compute conditional probabilities of various events of interest. Yet it is still challenging to analyze the "macroscopic" behavior of the order book, partly due to the strong coupling between buy-side and sell-side order flows. Computer simulation can also be slow because of the high-dimensionality of the Markov chain.

We address this issue by considering a scaling regime where the price tick size goes to zero and order flow rates tend to infinity. This regime is particularly relevant for high-frequency trading in the U.S. stock markets where stock price tick size is very small (one penny) and the duration of order book events is typically on the time scale of milliseconds. In this scaling regime, we develop a first-order fluid approximation to capture the sample path behavior of limit order book shape. Our main result (Theorem 6.1) states that a pair of measure-valued processes, representing the "empirical sell-side shape" and "empirical buy-side shape" of the order book, converges weakly to a pair of deterministic measure-valued processes in a certain Skorohod space. Moreover, the density profile of the limiting processes can be described by a simple Ordinary Differential Equation (ODE) whose coefficients are determined by the first-order statistics of the order flows.

Our contributions in this work are two-fold. First, we use the expected order flow

parameters to give a "macroscopic" description of the order book shape dynamics, whose order book event-level description is a multi-dimensional continuous-time Markov chain. Second, we perform experiments to test our theoretical model against order book data from NYSE Arca. The initial empirical results suggest that our model could potentially predict the order book shape evolution reasonably well for highly liquid stocks in a relatively stationary environment.

The approach in this work follows the martingale methods used in establishing the hydrodynamic limit for interacting particle systems, see, e.g., Kipnis and Landim [64], Liggett [72] and the references therein. The major challenge in our setting is the strong coupling between buy-side and sell-side order flows in limit order markets. In particular, in the model introduced in Cont et al. [21], the limit order arrival rates and cancellation rates on each price level depend on the distance to the opposite best quotes. It turns out in our scaling limit, the dynamics of buy and sell sides of the order book can be decoupled. This is achieved by assuming the rates of high-frequency order flows are "balanced" (Assumption 6.1), which leads to a separation of two time scales: a fast time scale for order book events, and a slow time scale for price changes due to depletion of volumes on best quotes. Thus in our hydrodynamic limit, we observe that the scaled best quotes remain unchanged while the rapidly varying order volumes on each price level are averaged out as a manifestation of the law of large numbers.

Similar scaling limits for two-sided order book shape dynamics have been established in Kruk [68], Bovier and Cerny [13], Kruk [69], Horsty and Paulsen [55], etc. However, our model, approach and result are all different from theirs. Kruk [68] considered a continuous auction model and proved a fluid limit for the number of orders on each of the finitely many price levels. Bovier and Cerny [13] proved a hydrodynamic limit of order book shape based on a certain particle reaction model. Their limit process is governed by a nonlinear parabolic Partial Differential Equation (PDE).

Kruk [69] utilized a similar order book model as in Roşu [92], and established a weak convergence result for a pair of measure-valued processes representing the buy-side and sell-side of the order book shape. Horsty and Paulsen [55] proved a law of large numbers result for the whole order book shape dynamics. In their limit, the bid and ask price dynamics can be described by two coupled ODEs and the relative buy and sell volume density functions can be described by two linear first-order hyperbolic PDEs.

There are several other related articles. Biais et al. [10], Bouchaud et al. [12], Foucault et al. [37], Potters and Bouchaud [84], and others studied statistical properties of order books on various financial markets. Russell and Kim [94] proposed a statistical forecasting model to capture the dependence in the order book shape. Lasry and Lions [70] proposed a mean-field model for the dynamical formation of price and evolution of order book shape. This model was further modified in Lehalle et al. [71] to replicate price volatility and the fast dynamics of real limit order books. Toth et al. [103] predicted a locally linear one-sided average order book shape and used it to explain the square-root price impact. Cont and De Larrard [20] proved a diffusion approximation for the volumes on the best quotes.

The rest of this chapter is organized as follows. Section 6.2 reviews the continuous-time Markov chain model in Cont et al. [21] for limit order book dynamics and states the assumptions on order flow rates and initial conditions. Section 6.3 summarizes our main result. Section 6.4 discusses empirical test of our result. Sections 6.5-6.7 are devoted to the proof of the main result. Auxiliary results are proved in the appendix.

6.1.1 Notation

This subsection contains all the notations used in this chapter. All random variables and stochastic processes are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ unless otherwise specified. Given $x \in \mathbb{R}$, we set $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. We

write $\lceil x \rceil$ for the smallest integer not less than x, and $\lfloor x \rfloor$ for the largest integer not greater than x. Given $x, y \in \mathbb{R}$, $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. For a positive integer n, $\mathbb{R}^n$ denotes the n-dimensional Euclidean space. The set of twice continuously differentiable functions on $\mathbb{S}$ is denoted by $C^2(\mathbb{S})$. The set of continuous functions on $[0, 1]$ is denoted by $\mathcal{C}([0, 1])$. Given a Polish space $\mathcal{E}$, the space of right-continuous functions $f : [0, T] \to \mathcal{E}$ with left limits is denoted by $\mathbb{D}([0, T], \mathcal{E})$. The space $\mathbb{D}([0, T], \mathcal{E})$ is assumed to be endowed with the Skorohod J_1-topology. For a sequence of random elements $\{X_n : n = 1, 2, \ldots\}$ taking values in a metric space, we write $X_n \Rightarrow X$ to denote the convergence of X_n to X in distribution. Each stochastic process with sample paths in $\mathbb{D}([0, T], \mathcal{E})$ is considered to be a $\mathbb{D}([0, T], \mathcal{E})$-valued random element. For a Borel measure ν and function f, we set $\langle \nu, f \rangle = \int f(u)\nu(du)$ when the integration exists. The symbol δ_u represents the Dirac measure at location $u \in [0, 1]$, i.e., for a Borel set U,

$$\delta_u(U) = \begin{cases} 1 & \text{if } u \in U, \\ 0 & \text{if } u \notin U. \end{cases}$$

The sign function is denoted by

$$sign(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

The space of finite *non-negative* measures with support contained in $[0, 1]$ is denoted by $\mathcal{M}^+([0, 1])$, and the space of finite *signed* measures with support contained in $[0, 1]$ is denoted by $\mathcal{M}([0, 1])$.

6.2 Model and Assumptions

In this section we describe the stochastic model introduced in Cont et al. [21] for order book dynamics and the assumptions on the order flow rates. Throughout, we

fix time $T > 0$.

Suppose that investors wish to submit their limit orders to n price levels $\{1, 2, \ldots, n\}$, which represent multiples of a price tick. The state of the limit order book are tracked by a continuous-time process $\mathcal{X}^n(t) \equiv (\mathcal{X}_1^n(t), \ldots, \mathcal{X}_n^n(t))$, where $|\mathcal{X}_i^n(t)|$ represents the number of outstanding limit orders at price $i \in \{1, 2, \ldots, n\}$ at time t. If $\mathcal{X}_i^n(t) > 0$, then there are $\mathcal{X}_i^n(t)$ sell orders at price i, and if $\mathcal{X}_i^n(t) < 0$, then there are $-\mathcal{X}_i^n(t)$ buy orders at price i. The best ask price $p_A^n(t)$ and best bid price $p_B^n(t)$ are defined by

$$p_A^n(t) = \inf\{i \in \{1, \ldots, n\} : \mathcal{X}_i^n(t) > 0\} \wedge n + 1, \tag{6.2.1}$$

$$p_B^n(t) = \sup\{i \in \{1, \ldots, n\} : \mathcal{X}_i^n(t) < 0\} \vee 0, \tag{6.2.2}$$

where $\inf \emptyset = \infty$ and $\sup \emptyset = -\infty$ by convention.

It is assumed that all the order flows are governed by independent Poisson processes. Specifically,

(a) Limit buy (respectively sell) orders arrive at a distance of i ticks from the opposite best quote at independent, exponentially distributed times with rate $\Lambda^n(i)$,

(b) Market buy (respectively sell) orders arrive at independent, exponentially distributed times with rate Υ^n,

(c) Each limit order at a distance of i ticks from the opposite best quote is cancelled independently after exponentially distributed times with rate $\Theta^n(i)$.

(d) The above events are mutually independent.

Given the above assumptions, the state process $\mathcal{X}^n(\cdot)$ is a $n-$dimensional continuous-time Markov chain and its infinitesimal generator $\mathcal{L}_n$ is given as follows: for any

$\mathcal{H} \in C^2(\mathbb{R}^n)$, and for each $u \in [0,T]$,

$$\begin{aligned}
\mathcal{L}_n\mathcal{H}(\mathcal{X}^n(u)) &= \sum_{k<p_A^n(u)} \left[\left(\mathcal{H}(\mathcal{X}^n(u)^{k+}) - \mathcal{H}(\mathcal{X}^n(u))\right)\cdot \Theta^n(p_A^n(u)-k)\cdot|\mathcal{X}_k^n(u)|\right] \\
&+ \sum_{k<p_A^n(u)} \left[\left(\mathcal{H}(\mathcal{X}^n(u)^{k-}) - \mathcal{H}(\mathcal{X}^n(u))\right)\cdot \Lambda^n(p_A^n(u)-k)\right] \\
&+ \sum_{k>p_B^n(u)} \left[\left(\mathcal{H}(\mathcal{X}^n(u)^{k+}) - \mathcal{H}(\mathcal{X}^n(u))\right)\cdot \Lambda^n(k-p_B^n(u))\right] \\
&+ \sum_{k>p_B^n(u)} \left[\left(\mathcal{H}(\mathcal{X}^n(u)^{k-}) - \mathcal{H}(\mathcal{X}^n(u))\right)\cdot \Theta^n(k-p_B^n(u))\cdot|\mathcal{X}_k^n(u)|\right] \\
&+ \left(\mathcal{H}(\mathcal{X}^n(u)^{p_B^n(u)^+}) - \mathcal{H}(\mathcal{X}^n(u))\right)\cdot \Upsilon^n \\
&+ \left(\mathcal{H}(\mathcal{X}^n(u)^{p_A^n(u)-}) - \mathcal{H}(\mathcal{X}^n(u))\right)\cdot \Upsilon^n, \qquad (6.2.3)
\end{aligned}$$

where for $k \in \{1,\dots,n\}$

$$\begin{cases}
\mathcal{X}^n(u)^{k+} = (\mathcal{X}_1^n(u),\dots,\mathcal{X}_{k-1}^n(u),\mathcal{X}_k^n(u)+1,\mathcal{X}_{k+1}^n(u),\dots,\mathcal{X}_n^n(u)), \\
\mathcal{X}^n(u)^{k-} = (\mathcal{X}_1^n(u),\dots,\mathcal{X}_{k-1}^n(u),\mathcal{X}_k^n(u)-1,\mathcal{X}_{k+1}^n(u),\dots,\mathcal{X}_n^n(u)).
\end{cases}$$

We now state our assumptions in this work. Our first assumption is on the high-frequency order flow rates.

Assumption 6.1 *There exist continuous functions $\Lambda(x)$ and $\Theta(x)$ on $[0,1]$ with $\Lambda(0)=\Theta(0)=0$, and constant Υ such that for each n,*

$$\frac{\Lambda^n(\lceil nx\rceil)}{n} = \Lambda(x), \quad \Theta^n(\lceil nx\rceil) = \Theta(x), \quad \textit{and} \quad \frac{\Upsilon^n}{n} = \Upsilon. \qquad (6.2.4)$$

Our second assumption concerns the initial order book shape.

Assumption 6.2 *There exist a continuous function ϱ from $[0,1]$ to $\mathbb{R}$ and a number $\mathsf{p} \in (0,1)$ such that*

$$\sup\{x\in[0,1] : \varrho(x) < 0\} = \inf\{x\in[0,1] : \varrho(x) > 0\} = \mathsf{p}. \qquad (6.2.5)$$

In addition, for fixed $n \geq 1$, the order book $\mathcal{X}^n = (\mathcal{X}_1^n,\dots,\mathcal{X}_n^n)$ is initialized according to the deterministic function ϱ:

$$\mathcal{X}_i^n(0) \equiv n\varrho(\frac{i}{n}). \qquad (6.2.6)$$

It is immediate from Assumption 6.2 and the definition of best bid and best ask prices that $\varrho(\mathsf{p}) = 0$, and

$$\lim_{n\to\infty} \frac{p_A^n(0)}{n} = \lim_{n\to\infty} \frac{p_B^n(0)}{n} = \mathsf{p}. \tag{6.2.7}$$

6.3 The main result

In this section we state our main result of this chapter. We first define three sequences of measure-valued process (see, e.g., Dawson [29] for background on measure-valued processes). For fixed $n \geq 1$ and $t \in [0, T]$, we set

$$\zeta_t^{n,+} = \frac{1}{n^2}\sum_{i=1}^{n}(\mathcal{X}_i^n(t))^+ \cdot \delta_{\frac{i}{n}} = \frac{1}{n^2}\sum_{i>p_B^n(t)} \mathcal{X}_i^n(t) \cdot \delta_{\frac{i}{n}}, \tag{6.3.1}$$

$$\zeta_t^{n,-} = \frac{1}{n^2}\sum_{i=1}^{n}(\mathcal{X}_i^n(t))^- \cdot \delta_{\frac{i}{n}} = \frac{1}{n^2}\sum_{i<p_A^n(t)} -\mathcal{X}_i^n(t) \cdot \delta_{\frac{i}{n}}, \tag{6.3.2}$$

$$\zeta_t^{n} = \zeta_t^{n,+} - \zeta_t^{n,-} = \frac{1}{n^2}\sum_{i=1}^{n} \mathcal{X}_i^n(t) \cdot \delta_{\frac{i}{n}}, \tag{6.3.3}$$

where $\mathcal{X}^n$ is the $n-$dimensional Markov chain with generator (6.2.3) describing the evolution of the limit order book, p_B^n, p_A^n are the best bid and best ask prices, and δ_u is the Dirac measure centered at u.

Note that $\zeta_\cdot^n$ is a Markov process and one can interpret ζ_t^n as the whole "empirical shape" of the order book at time t. However, ζ_t^n is a signed measure living in the space $\mathcal{M}([0,1])$, whose weak topology is known to be not metrizable (Varadarajan [104], Del Barrio and van de Geer [31]). Thus we instead work with the Markov process $(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-})$. For fixed t, the pair $(\zeta_t^{n,+}, \zeta_t^{n,-})$ represents the "empirical sell-side shape" and "empirical buy-side shape" of the order book at time t. This pair lives in the product space $\mathcal{M}^+([0,1]) \times \mathcal{M}^+([0,1])$. For the moment, we use $\bar{\mathcal{M}}([0,1])$ to denote this product space equipped with an appropriate metric such that $\bar{\mathcal{M}}([0,1])$ is complete and separable. A precise definition of $\bar{\mathcal{M}}([0,1])$ is given in Section 6.6.

We are interested in the limiting behavior of $(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-})$. When we send n to infinity, we have fast order flow rates by Assumption 6.1. One expects from this

assumption that the outstanding number of limit orders at each price level scales on the order of n. The choice of space scaling n^2 in (6.3.1), (6.3.2) and (6.3.3) then follows since there are n price levels in total. In addition, the price grids are rescaled from $\{1, \ldots, n\}$ to $\{\frac{1}{n}, \ldots, \frac{n}{n}\}$ in (6.3.1) and (6.3.2), therefore we are in the limiting regime of price tick size being 0 when we let $n \to \infty$. As mentioned earlier, this regime is particularly relevant for high-frequency trading in the U.S. stock markets where the duration between order book events is typically on the time scale of milliseconds, and the stock price tick size is very small (one penny).

The main result of this chapter is the following theorem. The proof is given in Sections 6.5-6.7.

Theorem 6.1 *Suppose that Assumptions 6.1 and 6.2 hold. Then as $n \to \infty$,*

$$(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) \Rightarrow (\zeta_\cdot^{+}, \zeta_\cdot^{-}) \quad \text{in } \mathbb{D}([0,T], \bar{\mathcal{M}}([0,1])), \tag{6.3.4}$$

where $(\zeta_\cdot^+, \zeta_\cdot^-)$ is a pair of deterministic measure-valued process. In addition, for any $t \in [0,T]$, the nonnegative measures ζ_t^+ and ζ_t^- are absolutely continuous with respect to Lebesgue measure and have density functions $\varphi^+(u,t)$ and $\varphi^-(u,t)$ such that

$$\begin{aligned} \zeta_t^{\pm}(du) &= \varphi^{\pm}(u,t)du, \\ \varphi^{\pm}(u,0) &= \varrho(u)^{\pm}, \qquad (6.3.5) \\ \partial_t \varphi^{\pm}(u,t) &= \Lambda((u-\mathsf{p})^{\pm}) - \Theta((u-\mathsf{p})^{\pm}) \cdot \varphi^{\pm}(u,t), \qquad (6.3.6) \end{aligned}$$

where Λ and Θ are functions given in Assumption 6.1, and ϱ and p are given in Assumption 6.2.

One readily verifies from (6.3.5) and (6.3.6) that

$$\begin{aligned} \varphi^+(u,t) &= e^{-\Theta((u-\mathsf{p})^+)t} \cdot \varrho(u)^+ + \frac{\Lambda((u-\mathsf{p})^+)}{\Theta((u-\mathsf{p})^+)}(1 - e^{-\Theta((u-\mathsf{p})^+)t}), \qquad (6.3.7) \\ \varphi^-(u,t) &= e^{-\Theta((u-\mathsf{p})^-)t} \cdot \varrho(u)^- + \frac{\Lambda((u-\mathsf{p})^-)}{\Theta((u-\mathsf{p})^-)}(1 - e^{-\Theta((u-\mathsf{p})^-)t}). \qquad (6.3.8) \end{aligned}$$

The function $\varphi^+(u,t)$ and $\varphi^-(u,t)$ represents the density profile of the sell-side and buy-side order book shape on the "macroscopic" time scale, i.e., one obtains from Theorem 6.1 the following approximation of the sample path behavior of order book shape: for $i \in \{1,\dots,n\}$

$$\frac{1}{n}\mathcal{X}_i^n(t) \approx \varphi^+(\frac{i}{n},t), \quad \text{if } \frac{i}{n} \geq \mathsf{p}, \tag{6.3.9}$$

$$\frac{1}{n}\mathcal{X}_i^n(t) \approx -\varphi^-(\frac{i}{n},t), \quad \text{if } \frac{i}{n} < \mathsf{p}. \tag{6.3.10}$$

6.4 Empirical test

In this section we test the sample path approximations of order book shape in (6.3.9) and (6.3.10) using order book data from NYSE Arca. As of 2009, NYSE Arca is the second largest electronic communication network in terms of shares traded. It accounts for approximately 20% of the trading volume for NASDAQ-listed securities and roughly 10% of NYSE-listed securities.

6.4.1 Data

Our data consists of one month of all limit order and market order activities on NYSE Arca in August 2010.

The limit order data contains three types of order action: add, delete and modify. "Add" corresponds to new limit order submission; "Delete" means that an order was cancelled, expired or filled; "Modify" signifies an order is modified either in its price, number of shares, or if an order is partially filled. The limit order data also contains a time stamp down to the millisecond, the price and order size, the buy or sell indicator, stock symbol, exchange, and an ID (identifier).

The market order data set records all the trades. It contains a time stamp down to the second, the traded price and order size, the buy or sell indicator, the best bid and ask prices when trades occur, stock symbol, and an ID (identifier).

The availability of these two sets of order flow message data enables us to reconstruct the limit order book at any give time for any stock traded on Arca. Moreover, one can analyze the limit order arrival rate, market order arrival rate and cancellation rate from those data sets.

6.4.2 Empirical test

In this section, we discuss empirical test of the sample path approximations of the order book shape in (6.3.9) and (6.3.10).

We focus on highly liquid stocks on which high frequency trading is prevalent and concentrated. The representative example we choose is Ford Motor Co. (symbol: F). The time period we study is 14:30 to 14:40 on August 16, 2010. We find from data that during this time window, the total number of order book events is 3697; the best bid price is \$12.28, and the best ask price is \$12.29.

We first estimate the order arrival rates Λ^n and cancel rates Θ^n during time window 14:30-14:40 as in Section 2 of Cont et al. [21]. Recall $\Lambda^n(i)$ and $\Theta^n(i)$ are the limit order arrival rate and limit order cancel rate at price i ticks away from the opposite best quote. The following two tables summarize the in-sample statistics of order flow intensities. The order flow rates at prices far away from the current best quotes are mostly zero; thus they are not displayed here.

Table 1: Limit order arrival rates 14:30-14:40 (unit: 100 shares/minute)

$\Lambda^n(1)$	$\Lambda^n(2)$	$\Lambda^n(3)$	$\Lambda^n(4)$	$\Lambda^n(5)$	$\Lambda^n(6)$	$\Lambda^n(7)$	$\Lambda^n(8)$	$\Lambda^n(9)$	$\Lambda^n(10)$
80	8	12	24	34	35	27	17	1	0

Table 2: Limit order cancel rates 14:30-14:40 (unit: 1/(100 shares · minute))

$\Theta^n(1)$	$\Theta^n(2)$	$\Theta^n(3)$	$\Theta^n(4)$	$\Theta^n(5)$	$\Theta^n(6)$	$\Theta^n(7)$	$\Theta^n(8)$	$\Theta^n(9)$	$\Theta^n(10)$
.6	.03	.06	.15	.28	.8	.3	.33	.04	0

We observe from the above two tables that the ratio $\frac{\Lambda^n(i)}{\Theta^n(i)}$ is on the scale of 100 for $i = 1, 2, \ldots, 9$. On the other hand, the natural choice of the scaling parameter n is 100 since the stock price tick size is $\$\frac{1}{100}$. Hence we have provided supporting evidence for Assumption 6.1.

Relying on those order flow estimates, we next test the model in (6.3.9). We begin with discussing the model inputs. The initial order book shape ϱ is given by linearly connecting the "volumes" (buy-side negative, and sell-side positive) on adjacent price levels at time 14:30. The number p is determined by the continuous function ϱ; p lies between best bid price \$12.28 and best ask price \$12.29. n is set to be 100. With these model inputs, one can compute φ^+ given in (6.3.7).

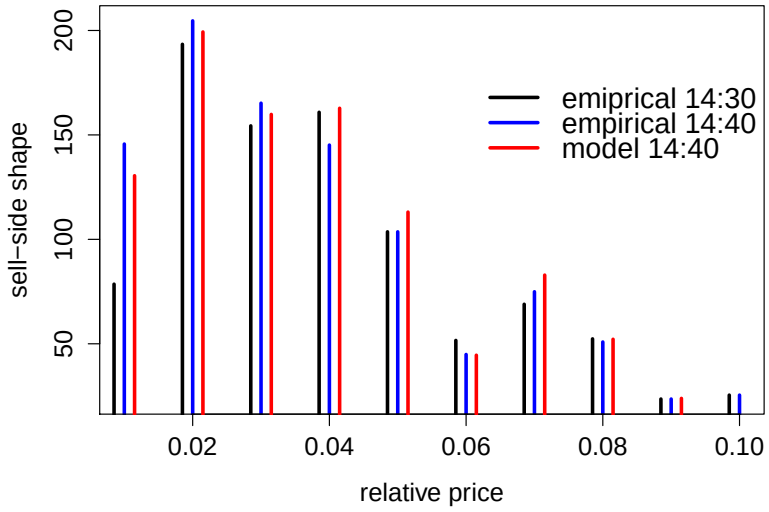

Figure 3: Stock F on Aug 16, 2010. The relative price represents the difference of limit sell price and best bid price. The model parameters (order flow rates) are estimated using data from 14:30 to 14:40.

Our result on testing of the model (6.3.9) is given in Figure 3. We observe good agreement of the model output and the empirical sell-side shape of stock F at 14:40.

Thus we have provided positive evidence that our model could yield a good approximation of the sample path behavior of order book shape on the "macroscopic" time scale.

For the purpose of prediction, we also estimated the order arrival rates in the time interval 14:20-14:30 for stock F on the same day. The following two tables summarize this out-of-sample statistics of order flow intensities. We again find that the ratio $\frac{\Lambda^n(i)}{\Theta^n(i)}$ is on the scale of 100 for $i = 1, \ldots, 8$. However, the order flow intensities are quite different from those in the time window 14:30-14:40.

Table 3: Limit order arrival rates 14:20-14:30 (unit: 100 shares/minute)

$\Lambda^n(1)$	$\Lambda^n(2)$	$\Lambda^n(3)$	$\Lambda^n(4)$	$\Lambda^n(5)$	$\Lambda^n(6)$	$\Lambda^n(7)$	$\Lambda^n(8)$	$\Lambda^n(9)$	$\Lambda^n(10)$
36.4	4.2	4.4	9.3	14	15.6	10	8.2	0	0

Table 4: Limit order cancel rates 14:20-14:30 (unit: $1/(100 \text{ shares} \cdot \text{minute})$)

$\Theta^n(1)$	$\Theta^n(2)$	$\Theta^n(3)$	$\Theta^n(4)$	$\Theta^n(5)$	$\Theta^n(6)$	$\Theta^n(7)$	$\Theta^n(8)$	$\Theta^n(9)$	$\Theta^n(10)$
.4	.04	.04	.067	.16	.2	.2	.14	.055	.6

Using these order flow rates, we perform an out-of-sample test of the model (6.3.9). Our result is given in Figure 4. We observe discrepancy of the model output and the empirical sell-side shape of stock F at 14:40. One major source of this discrepancy could be the non-stationarity of the order flow intensities.

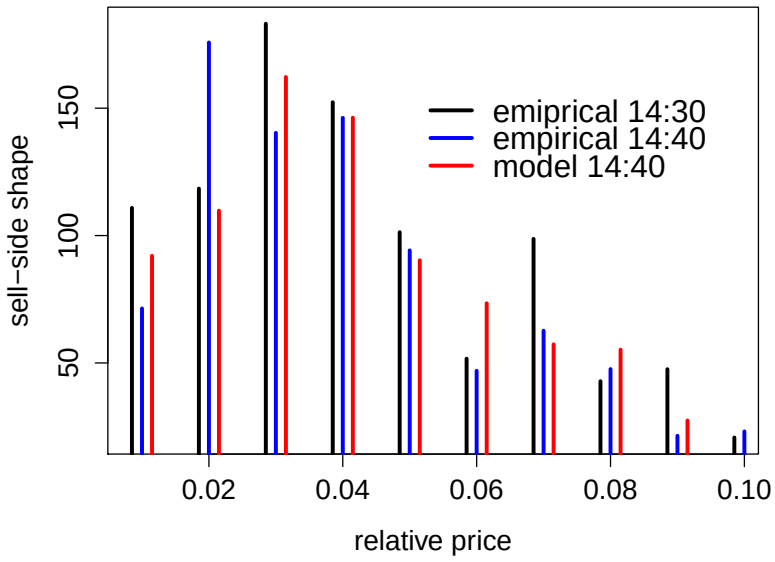

Figure 4: Stock F on Aug 16, 2010. The model parameters (order flow rates) are estimated using data from 14:20 to 14:30.

Therefore, we find from empirical tests that our model could potentially produce a good approximation of the evolution of order book shape on the "macroscopic" time scale. On the other hand, we remark that if one wants to predict the order book shape, more empirical tests need to be done here, including the analysis of other stocks and other time windows.

6.5 Convergence of best quotes

In this section we introduce a lemma which is critical for proving Theorem 6.1. It is concerned with the convergence of the "scaled" best bid and ask prices at any time $t \in [0, T]$. The proof is given in Appendix C.1.

Lemma 6.1 *Suppose Assumptions 6.1 and 6.2 hold. Then we have as* $n \to \infty$

$$\sup_{0\le t\le T}\left|\frac{p_A^n(t)}{n} - \mathsf{p}\right| \Rightarrow 0, \tag{6.5.1}$$

$$\sup_{0\le t\le T}\left|\frac{p_B^n(t)}{n} - \mathsf{p}\right| \Rightarrow 0, \tag{6.5.2}$$

where $p_A^n(t)$ *and* $p_B^n(t)$ *are given in (6.2.1) and (6.2.2), and* p *is given in (6.2.5).*

This lemma allows us to separate the dynamics of the buy and sell sides of the order book. This is in contrast with the model introduced in Section 6.2, where the buy-side and sell-side order flows are coupled through the best bid and best ask prices. We now explain the intuition for this lemma. From Assumption 6.1 one can expect that close to the best quotes there are thick queues on each price level with limit order volumes scale on the order of n. Given that the market order arrival rate is on the order of n, one expects that the duration of price changes due to the depletion of best quotes is on the order of 1. On the contrary, the duration between order book events is much shorter, which is on the order of $\frac{1}{n}$. Thus we have a separation of two time scales: a fast time scale for order book events, and a slow time scale for price changes due to depletion of volumes on best quotes. As a result, the volumes on each fixed price are averaged out but the "scaled" best quotes do not change in the hydrodynamic limit. On combining the initial condition of best quotes in (6.2.7), we obtain Lemma 6.1 intuitively.

6.6 Tightness of $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \ge 1\}$

In this section, we define the Polish space $\mathbb{D}([0,T], \bar{\mathcal{M}}([0,1]))$ on which the pair of measure-valued processes $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \ge 1\}$ lives and discuss the tightness of this sequence.

6.6.1 The Polish space $\mathbb{D}([0,T], \bar{\mathcal{M}}([0,1]))$

In this subsection, we define the Polish space $\bar{\mathcal{M}}([0,1]))$ and $\mathbb{D}([0,T], \bar{\mathcal{M}}([0,1]))$.

Recall that the space of finite *non-negative* measures with support contained in $[0,1]$, denoted by $\mathcal{M}^+([0,1])$, is a Polish space under the following metric d_+:

$$d_+(\upsilon_\alpha, \upsilon_\beta)) = \sum_{k=1}^{\infty} \frac{1}{2^k} \frac{|\langle \upsilon_\alpha, \phi_k\rangle - \langle \upsilon_\beta, \phi_k\rangle|}{1 + |\langle \upsilon_\alpha, \phi_k\rangle - \langle \upsilon_\beta, \phi_k\rangle|} \tag{6.6.1}$$

where $\upsilon_\alpha, \upsilon_\beta \in \mathcal{M}^+([0,1])$, $\{\phi_k : k \geq 1\}$ are chosen to be a dense subset of $\mathcal{C}([0,1])$, and $\langle \upsilon, \phi_k\rangle \equiv \int \phi_k(x)\upsilon(dx)$ for measure υ. The topology induced by metric d_+ is exactly the weak topology on $\mathcal{M}^+([0,1])$, i.e., $\upsilon_\alpha^n \Rightarrow \upsilon_\alpha$ if and only if $d_+(\upsilon_\alpha^n, \upsilon_\alpha) \to 0$ as $n \to \infty$. See, e.g, Kipnis and Landim [64, Section 4.1].

For fixed t and n, $(\zeta_t^{n,+}, \zeta_t^{n,-}) \in \mathcal{M}^+([0,1]) \times \mathcal{M}^+([0,1])$. As in Kotelenez [67], we define a metric d on this product space such that

$$d\left((\upsilon_\alpha^+, \upsilon_\alpha^-), (\upsilon_\beta^+, \upsilon_\beta^-)\right) \stackrel{\Delta}{=} \sqrt{d_+^2(\upsilon_\alpha^+, \upsilon_\beta^+) + d_+^2(\upsilon_\alpha^-, \upsilon_\beta^-)}, \tag{6.6.2}$$

where $(\upsilon_\alpha^+, \upsilon_\alpha^-), (\upsilon_\beta^+, \upsilon_\beta^-) \in \mathcal{M}^+([0,1]) \times \mathcal{M}^+([0,1])$ and the metric d_+ is give in (6.6.1). It is clear that

$$\max\{d_+(\upsilon_\alpha^+, \upsilon_\beta^+), d_+(\upsilon_\alpha^-, \upsilon_\beta^-)\} \leq d\left((\upsilon_\alpha^+, \upsilon_\alpha^-), (\upsilon_\beta^+, \upsilon_\beta^-)\right) \leq d_+(\upsilon_\alpha^+, \upsilon_\beta^+) + d_+(\upsilon_\alpha^-, \upsilon_\beta^-). \tag{6.6.3}$$

The product space $\mathcal{M}^+([0,1]) \times \mathcal{M}^+([0,1])$ equipped with metric d defined in (6.6.2) is denoted by $\bar{\mathcal{M}}([0,1])$. It follows from (6.6.3) and the fact that $\mathcal{M}^+([0,1])$ is a Polish space that $\bar{\mathcal{M}}([0,1])$ is also a Polish space.

Finally we discuss the Skorohod space $\mathbb{D}([0,T], \bar{\mathcal{M}}([0,1]))$ on which $(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-})$ lives. $\mathbb{D}([0,T], \bar{\mathcal{M}}([0,1]))$ is the space of paths mapping from $[0,T]$ to $\bar{\mathcal{M}}([0,1])$ that are right-continuous and have left limits everywhere, endowed with the Skorokhod J_1 topology. Since $\bar{\mathcal{M}}([0,1])$ is a Polish space with metric d, we deduce that the Skorohod space $\mathbb{D}([0,T], \bar{\mathcal{M}}([0,1]))$ is also a Polish space.

6.6.2 Tightness of $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$

We discuss the tightness of the sequence $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$ in this subsection. See Billingsley [11] for concepts and details on tightness in Skorohod space.

We first introduce a lemma, which reduces checking the tightness of the pair $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$ to checking the tightness of $\{\zeta_\cdot^{n,+} : n \geq 1\}$ and $\{\zeta_\cdot^{n,-} : n \geq 1\}$ individually. The proof directly follows from the tightness criteria in Lemma C.1 and the inequalities in (6.6.3), and thus is omitted here.

Lemma 6.2 *$\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$ is tight in $\mathbb{D}([0,T], \bar{\mathcal{M}}([0,1]))$ if and only if both $\{\zeta_\cdot^{n,+} : n \geq 1\}$ and $\{\zeta_\cdot^{n,-} : n \geq 1\}$ are tight in $\mathbb{D}([0,T], \mathcal{M}^+([0,1]))$.*

Now we state the key results in this subsection. Recall from (6.3.3) that we have for each $n \geq 1$, $t \in [0,T]$, and $f \in C^2([0,1])$,

$$\langle \zeta_t^n, f \rangle \quad \triangleq \quad \frac{1}{n^2} \sum_{i=1}^{n} \mathcal{X}_i^n(t) \cdot f(\frac{i}{n}). \tag{6.6.4}$$

Lemma 6.3 *Fix any $f \in C^2([0,1])$. The sequence of real-valued stochastic processes $\{\langle \zeta_\cdot^n, f \rangle : n \geq 1\}$ is tight in $\mathbb{D}([0,T], \mathbb{R})$.*

Combining Lemma 6.1 and Lemma 6.3, we can establish the tightness of the nonnegative measure-valued processes $\{\zeta_\cdot^{n,+} : n \geq 1\}$ and $\{\zeta_\cdot^{n,-} : n \geq 1\}$ as stated below.

Lemma 6.4 *$\{\zeta_\cdot^{n,+} : n \geq 1\}$ and $\{\zeta_\cdot^{n,-} : n \geq 1\}$ are both tight in $\mathbb{D}([0,T], \mathcal{M}^+([0,1]))$.*

We defer the proofs of these two lemmas to Appendices C.2 and *C*.3.

6.7 Limit points of $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$

In this section we characterize all the limit points of $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$. This sequence is tight and thus relatively compact by Prohorov's Theorem. Suppose $(\zeta_\cdot^+, \zeta_\cdot^-)$ is a limit point of this relatively compact sequence, i.e., there is a subsequence $\{\zeta_\cdot^{n_k} : k = 1, 2 \ldots\}$ such that

$$(\zeta_\cdot^{n_k,+}, \zeta_\cdot^{n_k,-}) \Rightarrow (\zeta_\cdot^+, \zeta_\cdot^-) \quad \text{in } \mathbb{D}([0,T], \bar{\mathcal{M}}([0,1])) \quad \text{as } n_k \to \infty \,. \tag{6.7.1}$$

We show in the next two subsections that $(\zeta_\cdot^+, \zeta_\cdot^-)$ are uniquely determined by some integrable equation.

6.7.1 The difference of the pair $(\zeta_\cdot^+, \zeta_\cdot^-)$

To characterize the pair $(\zeta_\cdot^+, \zeta_\cdot^-)$, we study their difference in this subsection. Define for each $t \in [0, T]$,

$$\zeta_t \stackrel{\Delta}{=} \zeta_t^+ - \zeta_t^-, \quad \text{and} \quad |\zeta_t| \stackrel{\Delta}{=} \zeta_t^+ + \zeta_t^-. \tag{6.7.2}$$

We first introduce a lemma on the absolute continuity of the measure $|\zeta_t|$ for fixed $t \in [0, T]$. This will be used in proving the uniqueness of the limit points of $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$ and identifying the limiting density profile. The proof is given in Appendix C.4.

Lemma 6.5 *For each $t \in [0, T]$, the measures $|\zeta_t|$ and ζ_t are absolutely continuous with respect to the Lebesgue measure.*

We need to introduce several additional notations to characterize $\zeta_\cdot$. Recall the functions Λ and Θ given in Assumption 6.1 and price p given in Assumption 6.2. Let ν_Λ be a signed measure absolutely continuous with respect to the Lebesgue measure and has density

$$\nu_\Lambda(dx) = \Lambda(|x - \mathsf{p}|) \cdot sign(x - \mathsf{p})dx, \quad \text{for } x \in [0, 1]. \tag{6.7.3}$$

Let $\mathcal{A}_\Theta$ be a linear operator such that for $f \in C^2([0, 1])$,

$$\mathcal{A}_\Theta f(x) = f(x)\Theta(|x - \mathsf{p}|). \tag{6.7.4}$$

We now state the central result in this subsection, which gives a characterization of $\zeta_\cdot$.

Lemma 6.6 *Let $\zeta_\cdot$ as defined in (6.7.2). Then $\zeta_\cdot$ is the unique deterministic signed-measure-valued process satisfying the following equation: for any $f \in C^2([0, 1])$ and $t \in [0, T]$,*

$$\langle \zeta_t, f \rangle = \langle \zeta_0, f \rangle + \langle \nu_\Lambda, f \rangle \cdot t - \int_0^t \langle \zeta_s, \mathcal{A}_\Theta f \rangle ds, \tag{6.7.5}$$

where ν_Λ is given in (6.7.3) and $\mathcal{A}_\Theta$ is given in (6.7.4).

Proof outline of Lemma 6.6 We use a martingale argument to establish (6.7.5). We critically rely on the following representation of the Markov process $\{\mathcal{X}^n(t) : t \geq 0\}$: for fixed $f \in C^2([0,1])$ and $n \geq 1$ we obtain from (6.6.4) that

$$\langle \zeta_t^n, f \rangle = \langle \zeta_0^n, f \rangle + \int_0^t \mathcal{L}_n F(\mathcal{X}^n(s)) ds + \mathsf{M}_t^n. \tag{6.7.6}$$

Here $\mathcal{L}_n$ is the generator operator for $\mathcal{X}^n$ as given in (6.2.3), M^n is a martingale with respect to the filtration generated by $\mathcal{X}^n$, and the function F is defined by

$$F(x_1, \ldots, x_n) = \frac{1}{n^2} \sum_{k=1}^{n} f(\frac{k}{n}) x_k. \tag{6.7.7}$$

We start with the weak converge of initial "empirical shape" of the order book.

Lemma 6.7 *Fix $f \in C^2([0,1])$. We have*

$$\lim_{n\to\infty} \langle \zeta_0^n, f \rangle = \langle \zeta_0, f \rangle = \int_0^1 f(x) \varrho(x) dx,$$

where ζ_0 is a deterministic signed measure absolutely continuous with respect to the Lebesgue measure with density $\zeta_0(dx) = \varrho(x)dx$ for $x \in [0,1]$.

The proof of Lemma 6.7 directly follows from (6.3.3) and (6.2.6), and is omitted.

The next lemma states that the martingale term in (6.7.6) vanishes as $n \to \infty$. The proof is provided in Appendix C.5.

Lemma 6.8

$$\sup_{0\leq t\leq T} |\mathsf{M}_t^n| \Rightarrow 0 \quad \text{as } n \to \infty. \tag{6.7.8}$$

Finally we consider the weak convergence of the generator $\{\mathcal{L}_n F(\mathcal{X}^n) : n \geq 1\}$. The proof is given in Appendix C.6.

Lemma 6.9 *For F given in (6.7.7) we have*

$$\sup_{0\leq s\leq T} |\mathcal{L}_n F(\mathcal{X}_s^n) + \langle \nu_\Lambda, f \rangle - \langle \zeta_s, \mathcal{A}_\Theta f \rangle| \Rightarrow 0 \quad \text{as } n \to \infty, \tag{6.7.9}$$

where ν_Λ is given in (6.7.3) and $\mathcal{A}_\Theta$ is given in (6.7.4).

With Lemma 6.7, Lemma 6.8 and Lemma 6.9 at our disposal, Equation (6.7.5) directly follows from (6.3.3), (6.7.2), (6.7.6) and the fact that $(\zeta_\cdot^+, \zeta_\cdot^-)$ is a limit point of the tight sequence $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$.

Next we turn to the proof of uniqueness of the solution of (6.7.5). Given the absolute continuity of ζ_t as stated in Lemma 6.5, we write $\zeta_t(du) = \varphi(u,t)du$ where $\varphi(u,t)$ is its density. We find from the integral equation (6.7.5) that φ is a weak solution for the following ODE system:

$$\varphi(u,0) = \varrho(u), \tag{6.7.10}$$

$$\partial_t\varphi(u,t) = \Lambda(|u-\mathsf{p}|)sign(u-\mathsf{p}) - \Theta(|u-\mathsf{p}|)\varphi(u,t). \tag{6.7.11}$$

On the other hand, it is clear that there is an unique classical solution to the ODE system (6.7.10) and (6.7.11). Therefore, we deduce from the equivalence of the classical solution and weak solution for ODEs that the density function φ uniquely solves (6.7.10) and (6.7.11). As a consequence, $\zeta_\cdot$ is the unique solution for (6.7.5). The proof is complete.

6.7.2 Hahn-Jordan decomposition

In this subsection, we show that the limiting pair $(\zeta_\cdot^+, \zeta_\cdot^-)$ is uniquely determined by their difference $\zeta_\cdot$. Our result is given below, whose proof is given in Appendix C.7. Recall the definition of ζ_t in (6.7.2).

Lemma 6.10 *$(\zeta_\cdot^+, \zeta_\cdot^-)$ is uniquely determined by $\zeta_\cdot$ in the sense that (ζ_t^+, ζ_t^-) is the Hahn-Jordan decomposition of the signed measure ζ_t for all $t \in [0,T]$.*

Proof of Theorem 6.1 Lemma 6.2 and Lemma 6.4 together imply that the sequence $\{(\zeta_\cdot^{n,+}, \zeta_\cdot^{n,-}) : n \geq 1\}$ is tight and thus it is relatively compact by Prohorov's Theorem. Suppose that some subsequence $\{(\zeta_\cdot^{n_k,+}, \zeta_\cdot^{n_k,-}) : k = 1, 2 \ldots\}$ weakly converges to a limit point $(\zeta_\cdot^+, \zeta_\cdot^-)$. Set $\zeta = \zeta^+ - \zeta^-$ as in (6.7.2). We deduce from Lemma 6.6 and Lemma 6.5 that the measure-valued process $\zeta_\cdot$ is deterministic and

its density function uniquely solves the ODE system in (6.7.10) and (6.7.11). Finally, Lemma 6.10 implies that the "one-dimensional" process $\zeta_\cdot$ uniquely determines the "two-dimensional" limiting process $(\zeta_\cdot^+, \zeta_\cdot^-)$ through Hahn-Jordan decomposition. We therefore deduce that the limit point $(\zeta_\cdot^+, \zeta_\cdot^-)$ is unique and the density functions of ζ_t^+ and ζ_t^- can be described by (6.3.5) and (6.3.6). The proof is complete.

APPENDIX A

APPENDIX FOR PART I

A.1 Proof of Proposition 3.3

We first outline the key idea behind the proof. Suppose that (B, g) is not controllable or that (B, h) is not observable in the CQLF existence problem. Then we can "reduce" them to suitable subspaces such that (B_1, g_1) is controllable and (B_1, h_1) is observable, where B_1 is a new matrix of lower dimension than B and similarly for g_1, h_1. In the process of "reduction", two desired properties are preserved: (a) $B(B - gh')$ has no real negative eigenvalues if and only if $B_1(B_1 - g_1h_1')$ has no real negative eigenvalues; (b) $(B, B - gh')$ has a CQLF if and only if $(B_1, B_1 - g_1h_1')$ has a CQLF. Therefore, applying Theorem 3.1 in Shorten et al. [95] to $(B_1, B_1 - g_1h_1')$ yields the result.

To make the ideas concrete, we now introduce a lemma giving an equivalent formulation of the CQLF existence problem, which makes the "reduction" possible. The lemma is an analog of Proposition 2 in King and Nathanson [63]. In King and Nathanson [63], each matrix of the pair is nonsingular while in our case one of the matrices is singular.

Lemma A.1 *Suppose that all eigenvalues of the matrix B have negative real part and all eigenvalues of $B - gh'$ have negative real part, except for a simple zero eigenvalue. Then the following statements are equivalent:*

(a) The pair $(B, B - gh')$ does not have a CQLF.

(b) There are positive semidefinite matrices X and Z such that

$$BX + XB' + (B - gh')Z + Z(B' - hg') = 0,$$

$$BX + XB' \neq 0 \quad \textit{and} \quad (B - gh')Z + Z(B' - hg') \neq 0.$$

(c) There are nonzero, positive semidefinite matrices X and Z such that

$$BX + XB' + (B - gh')Z + Z(B' - hg') = 0, \tag{A.1.1}$$

where $Z \neq cB^{-1}gg'(B^{-1})'$ for any $c \in \mathbb{R}$.

Proof We first prove the equivalence of (a) and (b). To set up the notation, let $S^{K\times K}$ be the space of real symmetric $K \times K$ matrices. For an arbitrary matrix $A \in \mathbb{R}^{K\times K}$, define the linear operator L_A on $S^{K\times K}$ by

$$L_A : S^{K\times K} \to S^{K\times K}, \quad L_A(H) = AH + HA'. \tag{A.1.2}$$

It is well-known that if A has eigenvalues $\{\lambda_i\}$ with eigenvectors $\{v_i\}$, then L_A has eigenvalues $\{\lambda_i + \lambda_j\}$ with eigenvectors $\{v_i v_j' + v_j v_i'\}$ for all $i \leq j$. Since all eigenvalues of the matrix B have negative real part, L_B is invertible.

Following King and Nathanson [63], we formulate the CQLF existence problem in terms of separating convex cones in $S^{K\times K}$. Define $\text{Cone}(B) = \{L_B(X) | X \geq 0\}$ and $\text{Cone}(B - gh') = \{L_{(B-gh')}(Z) | Z \geq 0\}$. Both are closed convex cones in $S^{K\times K}$. Let $S^{K\times K}$ be equipped with the usual Hilbert-Schmidt inner product $\langle X, Z\rangle = \text{tr}(XZ)$. We obtain that for any $Q \in S^{K\times K}$,

$$\langle X, QB + B'Q\rangle = \langle Q, BX + XB'\rangle = \langle Q, L_B(X)\rangle.$$

Note that for a nonzero positive semidefinite matrix X, we have $QB + B'Q < 0$ if and only if $\langle X, QB + B'Q\rangle < 0$, where the "if" part can be checked by taking $X = xx'$ for any nonzero $x \in \mathbb{R}^K$, and the "only if" part follows from the spectral decomposition of the positive semidefinite matrix X. Therefore, we have $QB + B'Q < 0$ if and only if $\langle Q, M\rangle < 0$ for all nonzero $M \in \text{Cone}(B)$. Using a similar argument one finds that $Q(B - gh') + (B - hg')Q \leq 0$ if and only if $\langle Q, T\rangle \leq 0$ for all nonzero $T \in \text{Cone}(B - gh')$. Moreover, since B only has eigenvalues with negative real part, we deduce that $QB + B'Q < 0$ for $Q \in S^{K\times K}$ implies that Q is positive

definite by Theorem 2.2.3 in Horn and Johnson [54]. By definition of CQLF, we thus obtain that $(B, B - gh')$ has a CQLF if and only if there exists a $Q \in S^{K\times K}$ such that $QB + B'Q < 0$ and $Q(B - gh') + (B - hg')Q \leq 0$. Equivalently, $(B, B - gh')$ has a CQLF if and only if there exists a $Q \in S^{K\times K}$ such that $\langle Q, M\rangle > 0$ for all nonzero $M \in \text{Cone}(-B)$ and $\langle Q, T\rangle \leq 0$ for all nonzero $T \in \text{Cone}(B - gh')$. Therefore, finding a CQLF for the pair $(B, B-gh')$ is the same as finding a separating hyperplane in $S^{K\times K}$ for $\text{Cone}(-B)$ and $\text{Cone}(B-gh')$. By the separating hyperplane theorem, we conclude that $(B, B-gh')$ not having a CQLF is equivalent to $\text{Cone}(-B)$ and $\text{Cone}(B - gh')$ having nonzero intersection. This completes the proof of the equivalence of (a) and (b).

We now turn to the equivalence of (b) and (c), for which we use the aforementioned spectral properties of the linear operator (A.1.2). Since L_B is invertible, we deduce that $L_B(X) = 0$ is equivalent to $X = 0$. We know that all eigenvalues of $(B-gh')$ have negative real part except for a simple zero eigenvalue, hence $L_{(B-gh')}$ also has a simple zero eigenvalue with eigenvector $cB^{-1}gg'(B^{-1})'$ for some nonzero $c \in \mathbb{R}$ while all of its other eigenvalues have negative real part. Consequently, $(B-gh')Z+Z(B-gh')' \neq 0$ is equivalent to $Z \neq cB^{-1}gg'(B^{-1})'$ for any $c \in \mathbb{R}$. The proof of the lemma is complete.

Proof of Proposition 3.3 In view of Theorem 3.1 of Shorten et al. [95], we need to check that controllability of (B, g) and observability of (B, h) need not be verified in the CQLF existence problem. Recall that controllability of (B, g) means that the vectors $g, Bg, B^2g, \ldots$ span $\mathbb{R}^K$, and observability of (B, h) means that the vectors $h, B'h, (B')^2h, \ldots$ span $\mathbb{R}^K$. To simplify the notation, let $\widetilde{B} = B - gh'$.

We first show that in the CQLF existence problem for the pair $(B, B - gh')$, we can assume without loss of generality that (B, g) is controllable. The proof relies on Lemma A.1. Let U be the span of vectors $g, Bg, B^2g \ldots$. Suppose U is a proper subspace of $\mathbb{R}^K$ with $\dim(U) < K$, and note that $\mathbb{R}^K = U \oplus U^\perp$ where $U^\perp$ is the orthogonal complement of U. In view of this decomposition, we perform a change of

basis and rewrite B, $\tilde{B}$, X and Y in the block form

$$B = \begin{pmatrix} B_1 & B_2 \\ 0 & B_3 \end{pmatrix}, \widetilde{B} = \begin{pmatrix} \widetilde{B}_1 & \widetilde{B}_2 \\ 0 & B_3 \end{pmatrix}, X = \begin{pmatrix} X_1 & X_2 \\ X_2' & X_3 \end{pmatrix}, Z = \begin{pmatrix} Z_1 & Z_2 \\ Z_2' & Z_3 \end{pmatrix}, \tag{A.1.3}$$

where $B - \widetilde{B} = gh'$ and g, h are represented in the new basis. We use the same notation for the matrices and vectors after the change of basis to save space, and we remark that the orthogonal transformation does not affect the existence of a CQLF for the pair $(B, \tilde{B})$ or the existence of real negative eigenvalues of $B\tilde{B}$. Namely, for any orthonormal matrix $O \in \mathbb{R}^K$, one readily checks that the pair $(B, \tilde{B})$ has a CQLF if and only if the pair $(OBO', O\tilde{B}O')$ has a CQLF. Furthermore, $B\tilde{B}$ has no real negative eigenvalues if and only if $(OBO')(O\tilde{B}O')$ has no real negative eigenvalues. Let g_1, h_1 be the orthogonal projection of g, h on the subspace U, so that $B_1 - \widetilde{B}_1 = g_1 h_1'$. Since U is the span of the vectors $g, Bg, B^2 g \ldots$, we deduce that $g_1, B_1 g_1, B_1^2 g_1 \ldots$ span U by (A.1.3), i.e., (B_1, g_1) is controllable. We now use Lemma A.1 to argue that there exists a CQLF for $(B, \widetilde{B})$ if and only if there exists a CQLF for $(B_1, \widetilde{B}_1)$, where (B_1, g_1) is controllable. Note that (A.1.3) implies, using (A.1.1) in Lemma A.1,

$$B_3(X_3 + Z_3) + (X_3 + Z_3)B_3' = 0.$$

Equivalently,

$$L_{B_3}(X_3 + Z_3) = 0,$$

where the linear operator L_{B_3} is defined in (A.1.2). Since B has only eigenvalues with negative real part, B_3 also has this property. This implies the linear operator L_{B_3} is invertible. We thus obtain $X_3 + Z_3 = 0$. Using the fact that X and Z are positive semidefinite, we deduce that $X_3 = Z_3 = 0$, and consequently $X_2 = Z_2 = 0$. This leads to

$$B_1 X_1 + X_1 B_1' + \widetilde{B}_1 Z_1 + Z_1 \widetilde{B}_1' = 0. \tag{A.1.4}$$

Thus for the pair $(B, B - gh')$, the existence of nonzero $X, Z \geq 0$ such that (A.1.1)

holds implies the existence of nonzero $X_1, Z_1 \geq 0$ such that (A.1.4) holds. Conversely, if there exists nonzero $X_1, Z_1 \geq 0$ such that (A.1.4) holds, setting $X_2 = X_3 = Z_2 = Z_3 = 0$, we then obtain that there exists nonzero $X, Z \geq 0$ such that (A.1.1) holds. Since $B - gh'$ has only eigenvalues with negative real part except for a simple zero eigenvalue, so does $B_1 - g_1 h_1'$. For $c \in \mathbb{R}$, since $g \in U$, one finds that $g'(B^{-1})' = (g_1'(B_1^{-1})', 0')$ by (A.1.3). Thus $Z \neq cB^{-1}gg'(B^{-1})'$ is equivalent to $Z_1 \neq cB_1^{-1}g_1 g_1'(B_1^{-1})'$. Putting these facts together, we apply Lemma A.1 to conclude that $(B, \widetilde{B})$ has no CQLF if and only if $(B_1, \widetilde{B}_1)$ has no CQLF, where (B_1, g_1) is controllable. Therefore, without loss of generality, we can assume that (B, g) is controllable in the CQLF existence problem for the pair $(B, B - gh')$.

We next show that without loss of generality we can assume that (B, h) is observable in the CQLF existence problem for the pair $(B, B - gh')$. Note that for $Q > 0$, we have $QB + B'Q < 0$ and $Q(B - gh') + (B' - hg')Q \leq 0$ if and only if $Q^{-1}B' + BQ^{-1} < 0$ and $Q^{-1}(B - hg') + (B' - gh')Q^{-1} \leq 0$. Hence $(B, B - gh')$ has a CQLF if and only if $(B', B' - hg')$ has a CQLF. From the preceding paragraph, we know that in the CQLF existence problem for the pair $(B', B' - hg')$, we can assume that (B', h) is controllable without loss of generality. By definition, (B', h) being controllable is the same as (B, h) being observable. Therefore, we conclude that we can assume without loss of generality that (B, h) is observable.

Finally, we argue that the pair $(B, B - gh')$ has a CQLF if and only if the matrix product $B(B - gh')$ has no real negative eigenvalues. Assuming that (B, g) is controllable and that (B, h) is observable, Theorem 3.1 in Shorten et al. [95] states that $(B, B - gh')$ has a CQLF if and only if the matrix product $B(B - gh')$ has no real negative eigenvalues. We have shown that we can always assume that (B, g) is controllable and that (B, h) is observable in the CQLF existence problem by reduction to proper subspaces. So it only remains to check that in the process of reduction,

the spectral property of having no real negative eigenvalues of the matrix product is preserved. Specifically, in the above proof that controllability of (B, g) can be assumed without loss of generality, we obtain that $(B, B - gh')$ has a CQLF if and only if $(B_1, B_1 - g_1 h_1')$ has a CQLF, where (B_1, g_1) is controllable. We next prove that $B(B - gh')$ has no real negative eigenvalues if and only if $B_1(B_1 - g_1 h_1')$ has no real negative eigenvalues, i.e., the desired spectral property of the matrix product is preserved in the process of reduction from $(B, B - gh')$ to $(B_1, B_1 - g_1 h_1')$. Observe that the spectrum of $B(B - gh')$ is the union of the spectrum of $B_1(B_1 - g_1 h_1')$ and B_3^2 by (A.1.3). Since all eigenvalues of B_3 have negative real part, we deduce that $B_1(B_1 - g_1 h_1')$ having no real negative eigenvalues is equivalent to $B(B - gh')$ having no real negative eigenvalues. A similar argument applies for observability instead of controllability. We have therefore completed the proof of Proposition 3.3.

A.2 Any quadratic function fails for $\hat{\alpha} > 0$

In this section, we give a simple example showing that, in general, no quadratic function can serve as a Lyapunov function in the Foster-Lyapunov criterion to prove positive recurrence of the piecewise OU process Y for $\hat{\alpha} > 0$. We first introduce a lemma which implies that the matrix $-\hat{R}(\mathsf{I} - \hat{p}e') - \hat{\alpha}\hat{p}e'$ is nonsingular for $\hat{\alpha} > 0$.

Lemma A.2 *If $\hat{\alpha} > 0$, then all eigenvalues of the matrix $-\hat{R}(\mathsf{I} - \hat{p}e') - \hat{\alpha}\hat{p}e'$ have negative real part.*

Proof It is clear that the matrix has an eigenvalue $-\hat{\alpha}$ with right eigenvector $\hat{p}$. Suppose $\hat{\lambda} \neq -\hat{\alpha}$ is an eigenvalue of the matrix with left eigenvector $\hat{\theta}$, i.e.,

$$\hat{\theta}'(-\hat{R}(\mathsf{I} - \hat{p}e') - \hat{\alpha}\hat{p}e') = \hat{\lambda}\hat{\theta}', \tag{A.2.1}$$

then we obtain that $\hat{\theta}'\hat{p} = 0$. It follows from (A.2.1) that $\hat{\lambda}$ is an eigenvalue of the matrix $-\hat{R}(\mathsf{I} - \hat{p}e')$. Moreover, $\hat{\lambda}$ cannot be zero since otherwise $\hat{\theta}' = ce'\hat{R}^{-1}$ for some nonzero $c \in \mathbb{R}$, which follows from the fact that $\hat{R}(\mathsf{I} - \hat{p}e')$ has a simple zero eigenvalue.

This contradicts the fact that $e'\hat{R}^{-1}\hat{p} > 0$ as seen in (3.4.30). From condition (b) in the proof of Theorem 3.1, we know that all nonzero eigenvalues of the matrix $-\hat{R}(\mathsf{I} - \hat{p}e')$ have negative real part. This completes the proof of the lemma.

Lemma A.3 *Suppose that Q is a real $K \times K$ positive semidefinite matrix such that at least one of the matrices $Q(-\hat{R}) + (-\hat{R}')Q$ and $Q(-\hat{R}(\mathsf{I} - \hat{p}e') - \hat{\alpha}\hat{p}e') + (-(\mathsf{I} - e\hat{p}')\hat{R}' - \hat{\alpha}e\hat{p}')Q$ fails to be negative definite. Let the quadratic function L be given by $L(y) = y'Qy$ for $y \in \mathbb{R}^K$. Then there exists some $\hat{\beta} \in \mathbb{R}$ and $v \in \mathbb{R}^K$ such that $GL(tv) \geq 0$ for any $t \geq 0$.*

Proof Suppose that $Q(-\hat{R}) + (-\hat{R}')Q$ fails to be negative definite, then there exists some $\lambda \geq 0$ and nonzero vector $v \in \mathbb{R}^K$ such that $[Q(-\hat{R}) + (-\hat{R}')Q]v = \lambda v$ and $e'v \leq 0$. By definition of the generator of Y, we thus obtain

$$\begin{aligned} GL(tv) &= \sum_{i,j} Q_{ij}(\sigma\sigma')_{ij} + (\nabla L(tv))'b(tv) \\ &= \sum_{i,j} Q_{ij}(\sigma\sigma')_{ij} + t^2 v'[Q(-\hat{R}) + (-\hat{R}')Q]v - 2t\hat{\beta}\hat{p}'Qv \\ &= \sum_{i,j} Q_{ij}(\sigma\sigma')_{ij} + \hat{\lambda}v'vt^2 - 2t\hat{\beta}\hat{p}'Qv. \end{aligned} \tag{A.2.2}$$

Since Q is positive semidefinite, we infer that $\sum_{i,j} Q_{ij}(\sigma\sigma')_{ij} = \mathrm{tr}(Q\sigma\sigma') = \mathrm{tr}(\sigma'Q\sigma) \geq 0$. Set $\hat{\beta} = 0$. We conclude from (A.2.2) that $GL(tv) \geq 0$ for any $t \geq 0$. A similar argument applies to the case where $Q(-\hat{R}(\mathsf{I} - \hat{p}e') - \hat{\alpha}\hat{p}e') + (-(\mathsf{I} - e\hat{p}')\hat{R}' - \hat{\alpha}e\hat{p}')Q$ fails to be negative definite. The proof of the lemma is complete.

In view of Lemmas A.2 and A.3, we give the following definition of strong CQLF which is slightly different from Definition 3.3 given in Section 3.2.1. For more details, refer to Shorten and Narendra [97] and King and Nathanson [63].

Definition A.1 (strong CQLF) *Let A and B be real $K \times K$ matrices having only eigenvalues with negative real part. For $Q \in \mathbb{R}^{K\times K}$, the quadratic form L given by*

$L(y) = y'Qy$ for $y \in \mathbb{R}^K$ is called a strong common quadratic Lyapunov function (strong CQLF) for the pair (A, B) if Q is positive definite and

$$QA + A'Q < 0,$$
$$QB + B'Q < 0.$$

We remark that it suffices to require Q to be a symmetric matrix in the above definition by Theorem 2.2.3 in Horn and Johnson [54].

We now formulate an example showing that, in general, no quadratic function can serve as a Lyapunov function in the Foster-Lyapunov criterion to prove positive recurrence of the piecewise OU process Y for $\hat{\alpha} > 0$. Let $\hat{R}$ be a matrix given by

$$\hat{R} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

so that $\hat{R}$ is a nonsingular M-matrix. Let $\hat{\alpha} = 133$ and $\hat{p}' = [0, 0, 1]$.

Lemma A.4 *For any quadratic function L given by $L(y) = y'Qy$ for some real $K \times K$ positive semidefinite matrix Q and all $y \in \mathbb{R}^K$, there exists some $\hat{\beta} \in \mathbb{R}$ and $v \in \mathbb{R}^K$ such that $GL(tv) \geq 0$ for any $t \in \mathbb{R}$ in the above example.*

Proof In view of Lemma A.3, it suffices to prove that there is no strong CQLF for the pair $(-\hat{R}, -\hat{R}(\mathsf{I} - \hat{p}e') - \hat{\alpha}\hat{p}e')$ for $\hat{\alpha} > 0$. Equivalently, it suffices to show that the matrix product $\hat{R}(\hat{R}(\mathsf{I} - \hat{p}e') + \hat{\alpha}\hat{p}e')$ has real negative eigenvalues by Theorem 1 in [63]. One readily checks that $\hat{R}(\hat{R}(\mathsf{I} - \hat{p}e') + \hat{\alpha}\hat{p}e')$ has three different eigenvalues: -7, $5 - \sqrt{82}$ and $5 + \sqrt{82}$. Thus, it has two real negative eigenvalues and we deduce that $(-\hat{R}, -\hat{R}(\mathsf{I} - \hat{p}e') - \hat{\alpha}\hat{p}e')$ has no strong CQLF in this example. Application of Lemma A.3 completes the proof of the lemma.

A.3 Positive Harris recurrence of $GI/Ph/n+M$ queues

In this appendix, we provide a sufficient condition on the interarrival time distribution under which Assumption 4.2 holds. To study the positive Harris recurrence of a $GI/Ph/n+M$ queue, we fix the number of servers n, as well as the interarrival distribution and phase-type service time distribution. Because n is fixed, unlike in Section 4.2, here we drop the superscript n in all relevant quantities. Let $\{\xi(i) : i = 0, 1, 2, \ldots\}$ be the sequence of interarrival times with $\{\xi(i) : i = 1, 2, \}$ being an i.i.d. sequence and $\xi(0)$ being the arrival time of the first customer after time 0. We use F to denote the cumulative distribution of $\xi(1)$. We make the following assumptions on F:

1. The distribution F is unbounded, i.e., $F(x) < 1$ for all $x > 0$.

2. The distribution F has a density, and furthermore the hazard function

$$h(x) = \frac{F'(x)}{1 - F(x)} \quad \text{for } x \geq 0$$

of the distribution F is locally bounded.

Recall that α is the rate of the exponential distribution for patience times. For the definition of positive Harris recurrence, see, for example, Dai [22]. In the following proposition, $Q(t)$ is the number of waiting customers in queue at time t. The other two quantities $A(t)$ and $Z(t)$ retain their meaning in Section 4.2.

Proposition A.1 *Suppose Assumptions (a)–(b) hold and $\alpha > 0$. The Markov process $\{(A(t), Q(t), Z(t)) : t \geq 0\}$ is positive Harris recurrent.*

Proof The process $\{(A(t), Q(t), Z(t)) : t \geq 0\}$ is known as a piecewise deterministic Markov process as defined in Davis [28]; see also Chapter 11 in Rolski et al. [91]. The main idea is to construct a suitable Lyapunov function f that is in the extended domain of the generator G of the Markov process. To construct the Lyapunov function,

we first define the mean residual life function $m(x)$ of the distribution function F by

$$m(x) = \frac{1}{1-F(x)} \cdot \int_x^\infty (1-F(s))ds = \mathbb{E}[\xi(1) - x | \xi(1) > x] \quad \text{for } x > 0.$$

We use $(a, q, z) = (a, q, z_1, \ldots, z_K) \in \mathcal{S} = \mathbb{R}_+ \times \mathbb{Z}_+ \times \mathbb{Z}_+^K$ to represent a state of the Markov process. For each state (a, q, z), we define

$$f(a, q, z) = 2F(a)(1 + m(a)) + q + \sum_{k=1}^{K} z_k. \tag{A.3.1}$$

The first component, $F(a)(1 + m(a))$, is identical to a Lyapunov function introduced in Lemma 2.1 of Konstantopoulos and Last [66]. We first verify that f is in the domain of the extended generator of the Markov process (see Definition 5.2 in Davis [28] for the definition of extended generator). We use Theorem 11.2.2 in Rolski et al. [91] (see also Theorem 5.5 in [28]) and check the three conditions there. Since the boundary set of the piecewise deterministic Markov process is empty, and the sample path of the age process A is absolutely continuous between jumps, it suffices to check

$$\mathbb{E}_{(a,q,z)} \left[\sum_{i:T_i \le t} |f(A(T_i), Q(T_i), Z(T_i)) - f(A(T_i-), Q(T_i-), Z(T_i-))| \right] < \infty, \tag{A.3.2}$$

where T_i are the jump epochs of the Markov process $\{(A(t), Q(t), Z(t)) : t \ge 0\}$. It follows from the proof of Lemma 2.1 and Equation (2.5) in Konstantopoulos and Last [66] that

$$\mathbb{E}_{(a,q,z)} \left[\sum_{i:T_i \le t} |F(A(T_i))(1 + m(A(T_i)) - F(A(T_i-))(1 + m(A(T_i-))| \right] < \infty.$$

Thus in order to establish (A.3.2), we know from (A.3.1) that it is enough to show

$$\mathbb{E}_{(a,q,z)} \left[\sum_{i:T_i \le t} \left| Q(T_i) + \sum_{k=1}^{K} Z_k(T_i) - Q(T_i-) - \sum_{k=1}^{K} Z_k(T_i-) \right| \right] < \infty. \tag{A.3.3}$$

To show (A.3.3), we note that the number of customer departures within $[0, t]$, due to either service completion or abandonment, is bounded by $q + n + E(t)$. Therefore,

$$\sum_{i:T_i \le t} \left| Q(T_i) + \sum_{k=1}^{K} Z_k(T_i) - Q(T_i-) - \sum_{k=1}^{K} Z_k(T_i-) \right| \le n + q + 2E(t),$$

from which (A.3.3) follows because

$$\mathbb{E}_{(a,q,z)}(E(t)) \leq \lambda t + C(\sqrt{t}+1) \quad \text{for some constant } C > 0.$$

See, e.g., Lemma 3.5 in Budhiraja and Ghosh [17].

Now we can write down the extended generator for the piecewise deterministic Markov process $\{(A(t), Q(t), Z(t)) : t \geq 0\}$. For $\sum_k z_k = n$, $q > 0$ and $a \geq 0$, it follows from (5.7) of Davis [28] that

$$\begin{aligned} Gf(a,q,z) &= \frac{\partial f}{\partial a}(a,q,z) + h(a) f(0,q+1,z) + \left(\alpha q + \sum_k \nu_k\right) f(a,q-1,z) \\ &\quad - (\alpha q + \sum_k \nu_k + h(a)) f(a,q,z) \\ &= 2\left[[F(a)(1+m(a))]' - h(a)[F(a)(1+m(a))]\right] + h(a) - \left(\alpha q + \sum_k \nu_k\right), \\ &\leq 2(1+h(a))\left[1 - 2F(a) + \int_a^\infty (1-F(s))ds\right] + h(a) - \left(\alpha q + \sum_k \nu_k\right), \end{aligned} \tag{A.3.4}$$

where in the second equality we have used the fact that $F(0) = 0$, and the inequality follows from the derivation on page 170 of Konstantopoulos and Last [66]. Because

$$\lim_{a\to\infty}\left[1 - 2F(a) + \int_a^\infty (1-F(s))ds\right] = -1,$$

there exists some $C_1 > 0$ such that

$$1 - 2F(a) + \int_a^\infty (1-F(s))ds \leq -\frac{1}{2} \quad \text{for } a > C_1. \tag{A.3.5}$$

Combining (6.2.3) and (A.3.5), we have for $a > C_1$ and $q > 0$

$$Gf(a,q,z) \leq -1 - \left(\alpha q + \sum_k \nu_k\right). \tag{A.3.6}$$

For any $a \geq 0$, $q = 0$ and $z \geq 0$, one can check that (A.3.6) continues to holds with $\sum_k \nu_k$ in (A.3.6) being replaced by $\sum_{k:z_k>0} \nu_k$. Therefore, we have for $a > C_1$, any $q \in \mathbb{Z}_+$, and any $z \in \mathbb{Z}_+^K$,

$$Gf(a,q,z) \leq -1. \tag{A.3.7}$$

Let

$$H = \sup_{0 \le a \le C_1} \left\{ 2(1+h(a))[1 - 2F(a) + \int_a^\infty (1 - F(s))ds] + h(a) \right\},$$

which is finite by Assumption (b) on F. It follows from (6.2.3) that for $a \in [0, C_1]$ and $q \ge C_2 = (H+1)/\alpha$, (A.3.7) also holds. It follows that

$$Gf(a, q, z) \le -1 + H 1_B(a, q, z), \tag{A.3.8}$$

where B is the compact set given by

$$B = \left\{ a \in [0, C_1],\ q \in [0, C_2] \cap \mathbb{Z}_+,\ z \in \mathbb{Z}_+^K,\ 0 \le \sum_k z_k \le n \right\}.$$

Since F is assumed to have density, it is spreadout. Together with Assumption (1) on F, this implies that the set B is a closed petite set in the state space $\mathcal{S} = \mathbb{R}_+ \times \mathbb{Z}_+ \times \mathbb{Z}_+^K$; see, e.g., the proof of Lemma 3.7 of Meyn and Down [75]. It follows from (A.3.8) and Theorem 4.2 of Meyn and Tweedie [77] that the Markov process $\{(A(t), Q(t), Z(t)) : t \ge 0\}$ is positive Harris recurrent.

A.4 Negative drift condition for the fluid model

It is the goal of this appendix to establish the negative drift condition for the fluid model, i.e., to prove Lemma 4.2. Throughout, we let $(\tilde{x}, \tilde{z})$ be defined through (4.4.7), i.e., as output of the map Ψ with input given in (4.4.6). The initial condition $(\tilde{x}(0), \tilde{z}(0)) \in \tilde{\mathcal{S}}$ is arbitrary.

We start with several auxiliary lemmas, which establish key properties of our Lyapunov function and the fluid model. Their proofs are deferred to the end of this appendix. The next lemma says that g is Lipschitz continuous.

Lemma A.5 *There exists a constant C such that*

$$\left| \sqrt{g(x^1, z^1)} - \sqrt{g(x^2, z^2)} \right| \le C |(x^1, z^1) - (x^2, z^2)|$$

for any $(x^1, z^1), (x^2, z^2) \in \tilde{\mathcal{S}}$.

The next lemma implies that $(\tilde{x}(t), \tilde{z}(t))$ has derivatives with respect to t almost everywhere.

Lemma A.6 *The function $t \mapsto (\tilde{x}(t), \tilde{z}(t))$ is locally Lipschitz continuous.*

We next formulate two important properties of the function g with fluid model input.

Lemma A.7 *Let $(\tilde{x}, \tilde{z})$ be the fluid model on $\tilde{\mathcal{S}}$ from Section 4.4.*

(a) For any $(\tilde{x}(0), \tilde{z}(0)) \in \tilde{\mathcal{S}}$, we have

$$\frac{dg(\tilde{x}(t), \tilde{z}(t))}{dt} \leq 0,$$

whenever $g(\tilde{x}(t), \tilde{z}(t))$ is differentiable at t.

(b) There are positive constants c, C and M such that, for any $(\tilde{x}(0), \tilde{z}(0)) \in \tilde{\mathcal{S}}$ such that $|(\tilde{x}(t), \tilde{z}(t))| \geq M$ and such that $g(\tilde{x}(t), \tilde{z}(t))$ is differentiable at t, we have

$$-C \cdot g(\tilde{x}(t), \tilde{z}(t)) \leq \frac{dg(\tilde{x}(t), \tilde{z}(t))}{dt} \leq -c \cdot g(\tilde{x}(t), \tilde{z}(t)).$$

With these three lemmas at our disposal, we are ready to prove the negative drift condition for the fluid model.

Proof of Lemma 4.2 Throughout this proof, when using constants from auxiliary lemmas, their subscript denotes the lemma they originated from.

Since $g(-\beta, -\beta\gamma) = 0$ or $g(-\mu\beta/\alpha, -\beta\gamma) = 0$, one obtains from Lemma A.5 that there exists some $C_{A.5}$ such that

$$\sqrt{g(x,z)} \leq C_{A.5}|(x,z)| + C_{A.5} \max\{|(\beta, \beta\gamma)|, |(\mu\beta/\alpha, \beta\gamma)|\}. \tag{A.4.1}$$

Let $c_{A.7}, C_{A.7}, M_{A.7}$ be the constants defined in Lemma A.7, and set

$$M = C_{A.5} M_{A.7} + C_{A.5} \max\{|(\beta, \beta\gamma)|, |(\mu\beta/\alpha, \beta\gamma)|\}.$$

It follows from (A.4.1) that if

$$\sqrt{g(\tilde{x}(t), \tilde{z}(t))} \geq M,$$

we have $|(\tilde{x}(t), \tilde{z}(t))| \geq M_{A.7}$ and thus by the second part of Lemma A.7,

$$-C_{A.7} \leq \frac{d}{dt} \ln g(\tilde{x}(t), \tilde{z}(t)) \leq -c_{A.7}. \tag{A.4.2}$$

Pick some C' such that

$$C' > M \cdot \exp\left(C_{A.7} t_0/2\right). \tag{A.4.3}$$

The constants ϵ and C from the statement of the lemma can be chosen as follows:

$$\epsilon = 1 - \exp\left(-\frac{1}{2} c_{A.7} t_0\right), \qquad C = \epsilon C',$$

as we now verify. If $\sqrt{g(x,z)} = \sqrt{g(\tilde{x}(0), \tilde{z}(0))} \leq C'$, then we have, by Lemma A.6 and the first part of Lemma A.7,

$$\sqrt{g(\tilde{x}(t_0), \tilde{z}(t_0))} - \sqrt{g(x,z)} \leq 0 = C - \epsilon C' \leq C - \epsilon \sqrt{g(x,z)}.$$

We next consider the case $\sqrt{g(x,z)} = \sqrt{g(\tilde{x}(0), \tilde{z}(0))} > C'$. We want to show that in this case

$$\sqrt{g(\tilde{x}(t), \tilde{z}(t))} \geq M \quad \text{for all } t \in [0, t_0], \tag{A.4.4}$$

which, together with (A.4.2), implies that

$$\sqrt{g(\tilde{x}(t_0), \tilde{z}(t_0))} \leq \exp(-c_{A.7} t_0/2) \sqrt{g(x,z)} = (1-\epsilon) \sqrt{g(x,z)}.$$

To establish (A.4.4), we assume that $\sqrt{g(\tilde{x}(0), \tilde{z}(0))} > C'$ and define

$$\tau = \inf\{t \geq 0 : \sqrt{g(\tilde{x}(t), \tilde{z}(t))} < M\}.$$

One readily checks that $\tau > 0$ and we now show that in fact $\tau > t_0$. If $\tau = \infty$, the claim is true. Now we assume $\tau < \infty$. For $t \in [0, \tau)$ we have

$$\sqrt{g(\tilde{x}(t), \tilde{z}(t))} \geq M,$$

where we have used the definition of τ. Next we can apply (A.4.2) for $t \in [0, \tau)$, and obtain

$$\ln g(\tilde{x}(0), \tilde{z}(0)) - C_{A.7}t \leq \ln g(\tilde{x}(t), \tilde{z}(t)) \leq \ln g(\tilde{x}(0), \tilde{z}(0)) - c_{A.7}t.$$

On combining this with the definition of τ, it follows that if $\sqrt{g(\tilde{x}(0), \tilde{z}(0))} > C'$,

$$\begin{aligned} \ln M = \ln \sqrt{g(\tilde{x}(\tau), \tilde{z}(\tau))} &\geq \ln \sqrt{g(\tilde{x}(0), \tilde{z}(0))} - C_{A.7}\tau/2 \\ &\geq \ln C' - C_{A.7}\tau/2 \\ &> \frac{1}{2}C_{A.7}(t_0 - \tau) + \ln M, \end{aligned}$$

where we use (A.4.3) in the last inequality. This shows that $\tau > t_0$, which proves (A.4.4) and therefore the statement of the lemma.

A.4.1 Proofs of auxiliary lemmas

We now prove Lemmas A.5, A.6, and A.7.

Proof of Lemma A.5 We first discuss the case $\beta \geq 0$. In that case, we have $g(x, z) = \|(x + \beta, z + \beta\gamma)\|_{\tilde{Q}}^2$, where

$$\|(x, z)\|_{\tilde{Q}}^2 = x^2 + \kappa z' \tilde{Q} z.$$

Note that $\|\cdot\|_{\tilde{Q}}$ defines a norm since $\tilde{Q}$ is positive definite.

Thus for $\beta \geq 0$, there exists a constant $C > 0$ such that

$$\begin{aligned} \left|\sqrt{g(x_1, z_1)} - \sqrt{g(x_2, z_2)}\right| &= |\|(x_1 + \beta, z_1 + \beta\gamma)\|_{\tilde{Q}} - \|(x_2 + \beta, z_2 + \beta\gamma)\|_{\tilde{Q}}| \\ &\leq \|(x_1 + \beta, z_1 + \beta\gamma) - (x_2 + \beta, z_2 + \beta\gamma)\|_{\tilde{Q}} \\ &= \|(x_1 - x_2, z_1 - z_2)\|_{\tilde{Q}} \\ &\leq C|(x_1, z_1) - (x_2, z_2)|, \end{aligned}$$

where the first inequality follows from the subadditivity property of the norm $\|\cdot\|_{\tilde{Q}}$, and the last inequality follows from the equivalence of the $|\cdot|$-norm and the $\|\cdot\|_{\tilde{Q}}$-norm

on the Euclidean space $\mathbb{R}^{K+1}$. We therefore obtain the claim when $\beta \geq 0$. The claim for $\beta < 0$ can be established similarly.

Proof of Lemma A.6 Since $t \mapsto (\tilde{u}(t), \tilde{v}(t))$ is continuous in t, we deduce from Lemma 9 in Dai et al. [24] that $t \mapsto (\tilde{x}(t), \tilde{z}(t))$ is also continuous in t. It follows from (4.4.4) that $t \mapsto \tilde{x}(t)$ is locally Lipschitz continuous. Moreover, using the fact that for all $s, t \geq 0$,

$$|\tilde{x}(t)^- - \tilde{x}(s)^-| \leq |\tilde{x}(t) - \tilde{x}(s)|,$$

we conclude from (4.4.5) that $\tilde{z}(\cdot)$ is also locally Lipschitz continuous, hence $(\tilde{x}(\cdot), \tilde{z}(\cdot))$ is locally Lipschitz.

For the proof of Lemma A.7, we need to study derivatives of the fluid model. From (4.4.4), (4.4.5), (4.4.6) and (4.4.7), it is straightforward to see that when $\tilde{x}(t) \geq 0$, we have $e'\tilde{z}(t) = 0$ and

$$\frac{d}{dt}\begin{bmatrix} \tilde{x}(t) \\ \tilde{z}(t) \end{bmatrix} = \begin{pmatrix} -\mu\beta \\ 0 \end{pmatrix} - \begin{pmatrix} \alpha & e'R \\ 0 & (\mathsf{I} - pe')R \end{pmatrix} \begin{bmatrix} \tilde{x}(t) \\ \tilde{z}(t) \end{bmatrix}. \tag{A.4.5}$$

On the other hand, when $\tilde{x}(t) < 0$, we have $e'\tilde{z}(t) = \tilde{x}(t)$, and

$$\frac{d}{dt}\begin{bmatrix} \tilde{x}(t) \\ \tilde{z}(t) \end{bmatrix} = \begin{pmatrix} -\mu\beta \\ -\mu\beta p \end{pmatrix} - \begin{pmatrix} 0 & e'R \\ 0 & R \end{pmatrix} \begin{bmatrix} \tilde{x}(t) \\ \tilde{z}(t) \end{bmatrix}. \tag{A.4.6}$$

We need two properties of the matrix $\tilde{Q}$ from (4.4.1) and (4.4.2), which are recorded in the following lemma.

Lemma A.8 *Let $\tilde{Q}$ be a positive definite matrix such that (4.4.1) and (4.4.2) hold.*

(a) The vectors $\tilde{Q}\gamma$ and e span the same one-dimensional space.

(b) Up to a multiplicative constant, γ is the only vector satisfying

$$[\tilde{Q}(\mathsf{I} - pe')R + R'(\mathsf{I} - ep')\tilde{Q}]\gamma = 0.$$

Proof One directly verifies that $\gamma = \mu R^{-1} p$ satisfies

$$\gamma'[\tilde{Q}(\mathsf{I} - pe')R + R'(\mathsf{I} - ep')\tilde{Q}]\gamma = 0.$$

Equation (4.4.2) states that $\tilde{Q}(\mathsf{I} - pe')R + R'(\mathsf{I} - ep')\tilde{Q}$ is a positive semi-definite matrix, and thus we deduce that

$$[\tilde{Q}(\mathsf{I} - pe')R + R'(\mathsf{I} - ep')\tilde{Q}]\gamma = 0,$$

which immediately implies that $\tilde{Q}(\mathsf{I} - pe')R\gamma = 0$ by the pdefinition of γ. Since $(\mathsf{I} - pe')$ has a simple zero eigenvalue and R is nonsingular, we must have

$$\tilde{Q}\gamma = be \quad \text{for some } b \neq 0,$$

as claimed in part (a).

As a corollary to part (a), we obtain $(\mathsf{I} - ep')\tilde{Q}R^{-1}p = 0$, which we use in the proof of part (b): it implies that

$$\begin{aligned}&\tilde{Q}(\mathsf{I} - pe')R + R'(\mathsf{I} - ep')\tilde{Q} \\ &\quad = R'(\mathsf{I} - ep') \cdot [(R^{-1})'\tilde{Q} + \tilde{Q}R^{-1}] \cdot (\mathsf{I} - pe')R.\end{aligned}$$

Equation (4.4.1) states that $QR + R'\tilde{Q}$ is positive definite. After left-multiplying by $(R^{-1})'$ and right-multiplying by R^{-1}, we find that $QR^{-1} + (R^{-1})'\tilde{Q}$ is also positive definite. Since $(\mathsf{I} - pe')R$ has rank $K - 1$ and has a simple zero eigenvalue with a right eigenvector $\gamma = \mu R^{-1} p$, we obtain from the preceding display that $\gamma = \mu R^{-1} p$ is the unique vector (up to a constant) such that

$$\gamma'[\tilde{Q}(\mathsf{I} - pe')R + R'(\mathsf{I} - ep')\tilde{Q}]\gamma = 0.$$

This implies part (b) of the lemma.

The next corollary controls the term involving z in our Lyapunov function.

Lemma A.9 *There exist constants $c, C > 0$ such that, whenever the derivative exists,*

$$-C|\tilde{z}(t)|^2 \leq \frac{d}{dt}\left[(\tilde{z}(t)+\beta\gamma)'\tilde{Q}(\tilde{z}(t)+\beta\gamma)\right] \leq -c|\tilde{z}(t)|^2,$$

as long as $e'\tilde{z}(t) = 0$ or equivalently $\tilde{x}(t) \geq 0$.

Proof It is readily seen with part (a) of Lemma A.8 that

$$\begin{aligned}\frac{d}{dt}\left[(\tilde{z}(t)+\beta\gamma)'\tilde{Q}(\tilde{z}(t)+\beta\gamma)\right] &= 2(\tilde{z}(t)+\beta\gamma)'\tilde{Q}\frac{d}{dt}\tilde{z}(t) \\ &= \tilde{z}(t)'[\tilde{Q}(\mathsf{I}-pe')R + R'(\mathsf{I}-ep')\tilde{Q}]\tilde{z}(t).\end{aligned}$$

Since $e'\gamma = 1 \neq 0$, we deduce from (4.4.2) and part (b) of Lemma A.8 that the quadratic form $h'[\tilde{Q}(\mathsf{I}-pe')R + R'(\mathsf{I}-ep')\tilde{Q}]h$ has a global (nonzero) minimum and maximum over the compact set $\{h \in \mathbb{R}^K : e'h = 0, |h| = 1\}$. This yields the statement of the lemma.

We are now ready to prove Lemma A.7.

Proof of Lemma A.7 We discuss the cases $\beta \geq 0$ and $\beta < 0$ separately.

Case 1: $\beta \geq 0$. In this case, we obtain from (4.4.3) that

$$\frac{d}{dt}g(\tilde{x}(t),\tilde{z}(t)) = \frac{d}{dt}(\tilde{x}(t)+\beta)^2 + \kappa\frac{d}{dt}\left[(\tilde{z}(t)+\beta\gamma)'\tilde{Q}(\tilde{z}(t)+\beta\gamma)\right]. \tag{A.4.7}$$

To compute the derivative with respect to t, we discuss two subcases.

Case 1.1: $\beta \geq 0$, $\tilde{x}(t) \geq 0$.

In this case we have $e'\tilde{z}(t) = 0$. We use the differential equation (A.4.5) to rewrite the expression in (A.4.7). For the first term, this yields

$$\frac{d}{dt}(\tilde{x}(t)+\beta)^2 = -2(\tilde{x}(t)+\beta)(\alpha\tilde{x}(t)+\mu\beta+e'R\tilde{z}(t)). \tag{A.4.8}$$

To bound this further, we use our assumption that β and $\tilde{x}(t)$ are nonnegative, which implies that

$$(\alpha\wedge\mu)(\tilde{x}(t)+\beta) \leq \mu\beta+\alpha\tilde{x}(t) \leq (\alpha\vee\mu)(\tilde{x}(t)+\beta),$$

where $\alpha \wedge \mu = \min\{\alpha, \mu\}$ and $\alpha \vee \mu = \max\{\alpha, \mu\}$.

We bound (A.4.8) from above as follows. Since $\alpha \wedge \mu > 0$, we deduce that there exists some (large) κ so that

$$\begin{aligned} |2(\tilde{x}(t)+\beta)e'R\tilde{z}(t)| &\leq 2|e'R| \cdot (\tilde{x}(t)+\beta) \cdot |\tilde{z}(t)| \\ &\leq (\alpha \wedge \mu)(\tilde{x}(t)+\beta)^2 + \kappa \frac{c_{A.9}}{2}|\tilde{z}(t)|^2, \end{aligned}$$

where $c_{A.9}$ is the constant from Lemma A.9. We have thus obtained

$$\frac{d}{dt}(\tilde{x}(t)+\beta)^2 \leq -(\alpha \wedge \mu)(\tilde{x}(t)+\beta)^2 + \kappa \frac{c_{A.9}}{2}|\tilde{z}(t)|^2.$$

Combining this with Lemma A.9 and (A.4.7), this yields

$$\frac{d}{dt} g(\tilde{x}(t), \tilde{z}(t)) \leq -(\alpha \wedge \mu)(\tilde{x}(t)+\beta)^2 - \kappa \frac{c_{A.9}}{2}|\tilde{z}(t)|^2,$$

which establishes part (a) of the statement in Case 1.1. It also gives the upper bound claimed in part (b), since the positive definiteness of $\tilde{Q}$ yields a constant $c' > 0$ such that

$$|\tilde{z}(t)|^2 \geq c'\tilde{z}(t)'\tilde{Q}\tilde{z}(t) \geq \frac{1}{2}c'(\tilde{z}(t)+\beta\gamma)'\tilde{Q}(\tilde{z}(t)+\beta\gamma),$$

where the last inequality holds outside of some compact set. To prove the lower bound claimed in part (b), we similarly note that

$$\begin{aligned} \frac{d}{dt}(\tilde{x}(t)+\beta)^2 &\geq -2(\alpha \vee \mu)(\tilde{x}(t)+\beta)^2 - 2(\tilde{x}(t)+\beta)e'R\tilde{z}(t) \\ &\geq -[2(\alpha \vee \mu) + (\alpha \wedge \mu)](\tilde{x}(t)+\beta)^2 - \kappa \frac{c_{A.9}}{2}|\tilde{z}(t)|^2, \end{aligned}$$

and one can bound the second term in (A.4.7) with Lemma A.9. We conclude that

$$\frac{d}{dt} g(\tilde{x}(t), \tilde{z}(t)) \geq -[2(\alpha \vee \mu) + (\alpha \wedge \mu)](\tilde{x}(t)+\beta)^2 - \kappa \left(\frac{c_{A.9}}{2} + C_{A.9}\right)|\tilde{z}(t)|^2.$$

By positive definiteness of $\tilde{Q}$ there exists some constant $C' > 0$ such that, outside of some compact set,

$$|\tilde{z}(t)|^2 \leq C'\tilde{z}(t)'\tilde{Q}\tilde{z}(t) \leq 2C'(\tilde{z}(t)+\beta\gamma)'\tilde{Q}(\tilde{z}(t)+\beta\gamma),$$

and we have thus shown all the claims in case $\beta \geq 0$ and $\tilde{x}(t) \geq 0$.

Case 1.2: $\beta \geq 0$, $\tilde{x}(t) < 0$.

In this case one has $e'\tilde{z}(t) = \tilde{x}(t)$. We use the differential equation (A.4.6) to rewrite (A.4.7). This leads to

$$\begin{aligned}
\frac{dg(\tilde{x}(t), \tilde{z}(t))}{dt} &= 2(\tilde{x}(t) + \beta)(-\mu\beta - e'R\tilde{z}(t)) \\
&\quad -\kappa(\tilde{z}(t) + \beta\gamma)'[\tilde{Q}R + R'\tilde{Q}](\tilde{z}(t) + \beta\gamma) \\
&= -2(\tilde{z}(t) + \beta\gamma)'ee'R(\tilde{z}(t) + \beta\gamma) \\
&\quad -\kappa(\tilde{z}(t) + \beta\gamma)'[\tilde{Q}R + R'\tilde{Q}](\tilde{z}(t) + \beta\gamma) \\
&= -(\tilde{z}(t) + \beta\gamma)'[ee'R + R'ee' + \kappa(\tilde{Q}R + R'\tilde{Q})](\tilde{z}(t) + \beta\gamma),
\end{aligned}$$

where we use $\gamma = \mu R^{-1}p$ and $\tilde{x}(t) + \beta = e'(\tilde{z}(t) + \beta\gamma)$. It follows from (4.4.1) that we can choose κ large so that $ee'R + R'ee' + \kappa(\tilde{Q}R + R'\tilde{Q})$ is positive definite. This immediately yields part (a) of the statement in case 1.2.

By definition of g, again using $e'\tilde{z}(t) = \tilde{x}(t)$, we also have

$$g(\tilde{x}(t), \tilde{z}(t)) = (\tilde{z}(t) + \beta\gamma)'[ee' + \kappa\tilde{Q}](\tilde{z}(t) + \beta\gamma).$$

Since $ee' + \kappa\tilde{Q}$ is also positive definite, the proof for the case $\beta \geq 0$ and $\tilde{x}(t) < 0$ is also complete.

Case 2: $\beta < 0$.

In this case, we obtain from (4.4.3) that

$$\frac{d}{dt}g(\tilde{x}(t), \tilde{z}(t)) = \frac{d}{dt}(\alpha\tilde{x}(t) + \mu\beta)^2 + \kappa\frac{d}{dt}\left[(\tilde{z}(t) + \beta\gamma)'\tilde{Q}(\tilde{z}(t) + \beta\gamma)\right]. \qquad \text{(A.4.9)}$$

As in Case 1, we discuss two subcases.

Case 2.1: $\beta < 0$, $\tilde{x}(t) \geq 0$. We use the differential equation (A.4.5) to rewrite the expression in (A.4.9). For the first term, this yields

$$\frac{d}{dt}(\alpha\tilde{x}(t) + \mu\beta)^2 = -2\alpha(\alpha\tilde{x}(t) + \mu\beta)(\alpha\tilde{x}(t) + \mu\beta + e'R\tilde{z}(t)).$$

The rest of the argument is almost identical to the one for Case 1.1. Increasing κ if necessary, one can show that

$$|2\alpha(\alpha\tilde{x}(t)+\mu\beta)e'R\tilde{z}(t)| \leq \alpha(\alpha\tilde{x}(t)+\mu\beta)^2 + \kappa\frac{c_{A.9}}{2}|\tilde{z}(t)|^2.$$

This leads to

$$\begin{aligned}
\frac{d}{dt}g(\tilde{x}(t),\tilde{z}(t)) &\leq -\alpha(\alpha\tilde{x}(t)+\mu\beta)^2 - \kappa\frac{c_{A.9}}{2}|\tilde{z}(t)|^2, \\
\frac{d}{dt}g(\tilde{x}(t),\tilde{z}(t)) &\geq -3\alpha(\alpha\tilde{x}(t)+\mu\beta)^2 - \kappa\left(\frac{c_{A.9}}{2}+C_{A.9}\right)|\tilde{z}(t)|^2.
\end{aligned}$$

The first inequality immediately establishes part (a) for $\beta < 0$, $\tilde{x}(t) \geq 0$. For part (b), one uses these two inequalities with the same arguments as in Case 1.1.

Case 2.2: $\beta < 0$, $\tilde{x}(t) < 0$. In this case we have $e'\tilde{z}(t) = \tilde{x}(t)$. We use the differential equation (A.4.6) to rewrite the expression in (A.4.9). This leads to

$$\begin{aligned}
\frac{d}{dt}g(\tilde{x}(t),\tilde{z}(t)) &= -2\alpha(\alpha\tilde{x}(t)+\mu\beta)(\mu\beta+e'R\tilde{z}(t)) \\
&\quad -\kappa(\tilde{z}(t)+\beta\gamma)'[\tilde{Q}R+R'\tilde{Q}](\tilde{z}(t)+\beta\gamma) \\
&= -2\alpha(\alpha\tilde{z}(t)+\mu\beta\gamma)'ee'R(\tilde{z}(t)+\beta\gamma) \\
&\quad -\kappa(\tilde{z}(t)+\beta\gamma)'[\tilde{Q}R+R'\tilde{Q}](\tilde{z}(t)+\beta\gamma),
\end{aligned}$$

where we use $\tilde{x}(t) = e'\tilde{z}(t)$. We next use an argument similar to the one used in Case 1.1. Since $\beta < 0$, $\tilde{z}(t)'e < 0$, we have

$$(\alpha\vee\mu)\cdot(\tilde{z}(t)+\beta\gamma)'e \leq (\alpha\tilde{z}(t)+\mu\beta\gamma)'e \leq (\alpha\wedge\mu)\cdot(\tilde{z}(t)+\beta\gamma)'e. \quad \text{(A.4.10)}$$

As a result, we obtain that

$$\begin{aligned}
|2\alpha(\alpha\tilde{x}(t)+\mu\beta)e'R(\tilde{z}(t)+\beta\gamma)| &\leq -2\alpha|e'R|\cdot|\tilde{z}(t)+\beta\gamma|(\alpha\tilde{z}(t)+\mu\beta\gamma)'e \\
&\leq 2\alpha(\alpha\vee\mu)|e'R|\cdot|\tilde{z}(t)+\beta\gamma|^2.
\end{aligned}$$

In view of (A.4.9), we therefore find that

$$\begin{aligned}
\frac{d}{dt}g(\tilde{x}(t),\tilde{z}(t)) &\leq 2\alpha(\alpha\vee\mu)|e'R|\cdot|\tilde{z}(t)+\beta\gamma|^2 - \kappa(\tilde{z}(t)+\beta\gamma)'[\tilde{Q}R+R'\tilde{Q}](\tilde{z}(t)+\beta\gamma) \\
\frac{d}{dt}g(\tilde{x}(t),\tilde{z}(t)) &\geq -2\alpha(\alpha\vee\mu)|e'R|\cdot|\tilde{z}(t)+\beta\gamma|^2 - \kappa(\tilde{z}(t)+\beta\gamma)'[\tilde{Q}R+R'\tilde{Q}](\tilde{z}(t)+\beta\gamma).
\end{aligned}$$

Increasing κ if necessary so that $-2\alpha(\alpha\vee\mu)|e'R|\mathsf{I} + \kappa[\tilde{Q}R + R'\tilde{Q}]$ is positive definite, we readily obtain part (a) of the lemma. For part (b), we need to make a comparison with g. By definition of g and (A.4.10), we obtain from $e'\tilde{z}(t) = \tilde{x}(t)$ that

$$\begin{aligned} g(\tilde{x}(t), \tilde{z}(t)) &\leq (\tilde{z}(t) + \beta\gamma)'[(\alpha \vee \mu)ee' + \kappa\tilde{Q}](\tilde{z}(t) + \beta\gamma) \\ g(\tilde{x}(t), \tilde{z}(t)) &\geq (\tilde{z}(t) + \beta\gamma)'[(\alpha \wedge \mu)ee' + \kappa\tilde{Q}](\tilde{z}(t) + \beta\gamma), \end{aligned}$$

which also yields part (b) of the claim in Case 2.2.

Combining the four cases, we have completed the proof of the lemma.

A.5 Negative drift condition for the diffusion-scaled processes

It is the goal of this appendix to establish the negative drift condition for the diffusion scaled processes, i.e., to prove Lemma 4.3. For this, we use the negative drift condition for the fluid model from Lemma 4.2.

Following Sections 4 and 5 in Dai et al. [24], we can use the map Ψ in Lemma 4.1 to represent the diffusion-scaled state processes given in (2.2.6) and (2.2.5):

$$(\tilde{X}^n, \tilde{Z}^n) = \Psi(\tilde{U}^n, \tilde{V}^n), \tag{A.5.1}$$

where the exact form of the processes $\tilde{U}^n$ and $\tilde{V}^n$ is not important at this point; they are specified in the proof of Lemma A.10 below. In view of this identity, we establish the negative drift condition for the diffusion-scaled processes by comparing the diffusion-scaled inputs $\tilde{U}^n$ and $\tilde{V}^n$ of the map Ψ with their fluid analogs $\tilde{u}^n$ and $\tilde{v}^n$, and then leveraging the negative drift condition of the fluid model.

Our negative drift result allows the diffusion-scaled process $(\tilde{A}^n, \tilde{X}^n, \tilde{Z}^n)$ to start from an arbitrary initial condition $(\tilde{A}^n(0), \tilde{X}^n(0), \tilde{Z}^n(0)) = (a, x, z) \in \tilde{\mathcal{S}}^n$. We initialize the fluid model with the same point (x, z), i.e., $(\tilde{x}(0), \tilde{z}(0)) = (x, z)$. As a result, the fluid model depends on n through the state space $\tilde{\mathcal{S}}^n$ of its initial point. Throughout this appendix, we stress this dependence by writing $(\tilde{x}^n, \tilde{z}^n)$ for the fluid

model instead of $(\tilde{x}, \tilde{z})$. Similarly, we write $\tilde{u}^n$ and $\tilde{v}^n$ instead of $\tilde{u}$ and $\tilde{v}$, as defined through (4.4.6).

The following auxiliary lemma ensures that the inputs to Ψ are close to their fluid analogs. Its proof is deferred to the end of this appendix.

Lemma A.10 *Fix $t_0 \geq 0$. There exists a constant $C = C(t_0) > 0$ such that for each n large enough and each initial state $(a, x, z) = (\tilde{a}^n(0), \tilde{x}^n(0), \tilde{z}^n(0)) \in \tilde{\mathcal{S}}^n$, we have*

$$\mathbb{E}_{(a,x,z)}\left[\|\tilde{U}^n - \tilde{u}^n\|_{t_0}\right] < C + C\sqrt[4]{g(x,z)}. \tag{A.5.2}$$

$$\mathbb{E}_{(a,x,z)}\left[\|\tilde{V}^n - \tilde{v}^n\|_{t_0}\right] < C + C\sqrt[4]{g(x,z)}. \tag{A.5.3}$$

With Lemma A.10 at our disposal, we are ready to prove the negative drift condition for diffusion-scaled processes.

Proof of Lemma 4.3 Throughout this proof, when using constants from auxiliary lemmas, their subscript again denotes the lemma from which they originated.

Let n be large enough as in Lemma A.10 and let $(a, x, z) = (\tilde{a}^n(0), \tilde{x}^n(0), \tilde{z}^n(0)) \in \tilde{\mathcal{S}}^n$. We first note that

$$\begin{aligned}
&\left|\sqrt{g(\tilde{X}^n(t_0), \tilde{Z}^n(t_0))} - \sqrt{g(\tilde{x}^n(t_0), \tilde{z}^n(t_0))}\right| \\
&\leq C_{A.5}|(\tilde{X}^n(t_0), \tilde{Z}^n(t_0)) - (\tilde{x}^n(t_0), \tilde{z}^n(t_0))| \\
&= C_{A.5}|\Psi(\tilde{U}^n, \tilde{V}^n)(t_0) - \Psi(\tilde{u}^n, \tilde{v}^n)(t_0)| \\
&\leq C_{A.5}\|\Psi(\tilde{U}^n, \tilde{V}^n) - \Psi(\tilde{u}^n, \tilde{v}^n)\|_{t_0} \\
&\leq C_{4.1}(t_0)C_{A.5}[\|\tilde{U}^n - \tilde{u}^n\|_{t_0} + \|\tilde{V}^n - \tilde{v}^n\|_{t_0}],
\end{aligned} \tag{A.5.4}$$

where the first inequality follows from Lemma A.5, the equality follows from the fact that $(\tilde{X}^n, \tilde{Z}^n) = \Psi(\tilde{U}^n, \tilde{V}^n)$ and $(\tilde{x}^n, \tilde{z}^n) = \Psi(\tilde{u}^n, \tilde{v}^n)$, and the last inequality follows from the Lipschitz continuity of the map Ψ as in part (c) of Lemma 4.1.

It follows from (A.5.4), Lemma 4.2, and Lemma A.10 that

$$\begin{aligned}
&\mathbb{E}_{(a,x,z)}\left(\sqrt{g(\tilde{X}^n(t_0),\tilde{Z}^n(t_0))}-\sqrt{g(x,z)}\right)\\
&\leq\ \mathbb{E}_{(a,x,z)}\left|\sqrt{g(\tilde{X}^n(t_0),\tilde{Z}^n(t_0))}-\sqrt{g(\tilde{x}^n(t_0),\tilde{z}^n(t_0))}\right|\\
&\quad+\left(\sqrt{g(\tilde{x}^n(t_0),\tilde{z}^n(t_0))}-\sqrt{g(x,z)}\right)\\
&\leq\ 2C_{4.1}(t_0)C_{A.5}C_{A.10}(1+\sqrt[4]{g(x,z)})+C_{4.2}-\epsilon_{4.2}\sqrt{g(x,z)}. \qquad \text{(A.5.5)}
\end{aligned}$$

Pick some $C'>0$ such that, for any $b\in\mathbb{R}$ with $b\geq C'$,

$$2C_{4.1}(t_0)C_{A.5}C_{A.10}\sqrt[4]{b}-\frac{1}{2}\epsilon_{4.2}\sqrt{b}\leq 0.$$

The constants C and ϵ in the statement of the lemma are chosen as follows:

$$C=2C_{4.1}(t_0)C_{A.5}C_{A.10}(1+\sqrt[4]{C'})+C_{4.2},\qquad \epsilon=\frac{1}{2}\epsilon_{4.2},$$

and we now verify the statement of the lemma with these definitions. If (x,z) satisfies $g(x,z)<C'$, then the right-hand side of (A.5.5) is bounded from above by

$$2C_{4.1}(t_0)C_{A.5}C_{A.10}(1+\sqrt[4]{C'})+C_{4.2}-\epsilon_{4.2}\sqrt{g(x,z)}=C-2\epsilon\sqrt{g(x,z)}.$$

If (x,z) satisfies $g(x,z)\geq C'$, then it is bounded from above by

$$2C_{4.1}(t_0)C_{A.5}C_{A.10}+C_{4.2}-\frac{1}{2}\epsilon_{4.2}\sqrt{g}=C-\epsilon\sqrt{g(x,z)}-C_{4.1}(t_0)C_{A.5}C_{A.10}\sqrt[4]{C'}.$$

We have thus obtained (4.5.1) in both cases.

A.5.1 Proof of auxiliary lemma

We now prove Lemma A.10.

Proof of Lemma A.10 In this proof, the constant C is a generic constant independent of n, but may vary line from line. Fix $t_0>0$. We start by specifying the processes $\tilde{U}^n$ and $\tilde{V}^n$ for which (A.5.1) holds. Following (5.10) and (5.11) in Dai et

al. [24], we have

$$\tilde{U}^n(t) = \tilde{X}^n(0) - \mu\beta^n t + \tilde{E}^n(t) + e'\tilde{M}^n(t) - \tilde{G}^n\left(\int_0^t (\bar{X}^n(s))^+ \, ds\right), \tag{A.5.6}$$

$$\tilde{V}^n(t) = (\mathsf{I} - pe')\tilde{Z}^n(0) + \tilde{\Phi}^{0,n}(\bar{B}^n(t)) + (\mathsf{I} - pe')\tilde{M}^n(t). \tag{A.5.7}$$

The processes $\tilde{E}^n, \tilde{M}^n, \tilde{G}^n, \tilde{\Phi}^{0,n}$, $\bar{X}^n$, and $\bar{B}^n$ are defined as follows, see Section 5.2 of Dai et al. [24]. First, $\tilde{E}^n$ is given in (2.2.2). For each $k = 1, \ldots, K$, let S_k be a Poisson process with rate ν_k, and let G be a Poisson process with rate α. For each $n \geq 1$ define the diffusion-scaled processes:

$$\tilde{S}_k^n(t) = \frac{1}{\sqrt{n}}(S_k(nt) - n\nu_k t), \quad \tilde{G}^n(t) = \frac{1}{\sqrt{n}}(G(nt) - n\alpha t) \quad t \geq 0. \tag{A.5.8}$$

Moreover, for each $N \geq 1$ and $k = 0, \ldots, K$, define the routing processes

$$\Phi^k(N) = \sum_{j=1}^{N} \phi^k(j),$$

where $\{\phi^k(j) : j = 1, 2, \ldots\}$ are i.i.d. Bernoulli random vectors with mean p^k. Here $p^0 = p$, and p^k is the kth column of matrix P'. For each $n \geq 1$, set

$$\tilde{\Phi}^{k,n}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (\phi^k(j) - p^k) \quad t \geq 0, \tag{A.5.9}$$

where, for an $x \in \mathbb{R}$, $\lfloor x \rfloor$ is the largest integer that is less than or equal to x. Let $T_k^n(t)$ be the cumulative amount of service effort received by customers in phase k service in $(0, t]$, $B^n(t)$ be the cumulative number of customers who have entered into service by time t. Then $\tilde{M}^n(t)$ is defined via

$$\tilde{M}^n(t) = \sum_{k=1}^{K} \tilde{\Phi}^{k,n}(\bar{S}_k^n(\bar{T}_k^n(t))) - (\mathsf{I} - P')\tilde{S}^n(\bar{T}^n(t)), \tag{A.5.10}$$

where $\bar{S}^n(t) = S(nt)/n$ and $\bar{T}^n(t) = T_n(t)/n$ for $t \geq 0$. We also let $\bar{B}^n(t) = B^n(t)/n$ for $t \geq 0$. Finally, we have

$$\bar{X}^n(t) = \frac{1}{n} X^n(t),$$

where $X^n(t)$ is given in (2.2.3). As in Dai et al. [24], we have that

$$\tilde{E}^n, \tilde{S}_1^n, \ldots, \tilde{S}_K^n, \tilde{\Phi}^{0,n}, \ldots, \tilde{\Phi}^{K,n} \text{ and } \tilde{G}^n \text{ are mutually independent.}$$

Having explained the meaning of all processes involved, we now proceed and prove the statement of Lemma A.10. By (A.5.6), (A.5.7), and (4.4.6) we have

$$\|\tilde{U}^n - \tilde{u}^n\|_{t_0} \leq \mu|\beta^n - \beta|t_0 + \|\tilde{E}^n\|_{t_0} + \sqrt{K}\|\tilde{M}^n\|_{t_0} + \left\|\left|\tilde{G}^n\left(\int_0^t (\bar{X}^n(s))^+ \, ds\right)\right|\right\|_{t_0}, \tag{A.5.11}$$

$$\|\tilde{V}^n - \tilde{v}^n\|_{t_0} \leq \|\tilde{\Phi}^{0,n}(\bar{B}^n)\|_{t_0} + (1 + |p|\sqrt{K})\|\tilde{M}^n\|_{t_0}. \tag{A.5.12}$$

We first establish (A.5.2). The proof is similar to that of Lemma 3.5 in Budhiraja and Ghosh [17] and we only highlight the main differences. We start with inequality (A.5.11). We bound the three terms in the right side of (A.5.11) separately. Let $\{E^*(t) : t \geq 0\}$ be a renewal process associated with i.i.d. sequence $\{u(i) : i \geq 1\}$ that was given in Assumption 4.1, where $u(1)$ is the first renewal time. It follows from Assumption 4.1 that for each $t \geq 0$,

$$E^*(\lambda^n t) \leq E^n(t) \leq 1 + E^*(\lambda^n t). \tag{A.5.13}$$

Since $E^*(\cdot)$ is independent of $(A^n(0), X^n(0), Z^n(0))$ for any n, using the definition of $\tilde{E}^n$ in (2.2.2) and Equation (A.5.13), we deduce that there exists a constant C independent of n and of the initial states (a, x, z) such that

$$\begin{aligned}
\mathbb{E}_{(a,x,z)}\|\tilde{E}^n\|_{t_0}^2 &\leq \mathbb{E}_{(a,x,z)} \sup_{0\leq t\leq t_0} \frac{1}{n} \max\{|1 + E^*(\lambda^n t) - \lambda^n t|^2, |E^*(\lambda^n t) - \lambda^n t|^2\} \\
&\leq \mathbb{E}_{(a,x,z)} \sup_{0\leq t\leq t_0} \frac{2}{n}(|E^*(\lambda^n t) - \lambda^n t|^2 + 1) \\
&= \mathbb{E} \sup_{0\leq t\leq t_0} \frac{2}{n}(|E^*(\lambda^n t) - \lambda^n t|^2 + 1) \\
&= 2\mathbb{E} \sup_{0\leq t\leq t_0} \frac{1}{\lambda^n}|E^*(\lambda^n t) - \lambda^n t|^2 \cdot \frac{\lambda^n}{n} + \frac{2}{n} \\
&\leq (t_0 + 1)C,
\end{aligned}$$

where the last inequality follows from Lemma 3.5 in Budhiraja and Ghosh [17] and the fact that the sequence $\{\lambda^n/n : n \geq 1\}$ is bounded. Hence we have

$$\mathbb{E}_{(a,x,z)}\|\tilde{E}^n\|_{t_0} \leq C + C\sqrt{t_0}. \tag{A.5.14}$$

Moreover, we know from (A.5.10) and Lemma 3.5 in [17] that

$$\mathbb{E}_{(a,x,z)}\|\tilde{M}^n\|_{t_0}^2 \leq (t_0+1)C,$$

where we use the fact that for each k, $\bar{T}_k^n(t) \leq t$ for all $t \geq 0$. This immediately implies

$$\mathbb{E}_{(a,x,z)}\|\tilde{M}^n\|_{t_0} \leq C + C\sqrt{t_0}. \tag{A.5.15}$$

Finally we show there is a constant $C(t_0)$ which depends on t_0 but is independent of n and any initial state $(a,x,z)\tilde{\mathcal{S}}^n$ such that

$$\mathbb{E}_{(a,x,z)}\left\|\tilde{G}^n\left(\int_0^t (\bar{X}^n(s))^+\,ds\right)\right\|_{t_0} \leq C(t_0) + C(t_0)\frac{\sqrt[4]{g(x,z)}}{\sqrt{n}}. \tag{A.5.16}$$

To see this, we know from (2.2.3) that for all $0 \leq s \leq t$

$$(\bar{X}^n(s))^+ \leq (\bar{X}^n(0))^+ + \bar{E}^n(t) = x^+/\sqrt{n} + \bar{E}^n(t),$$

where $x = \tilde{X}^n(0) = \sqrt{n}\bar{X}^n(0)$ and $\bar{E}^n(t) = \frac{1}{n}E^n(t)$ for $t \geq 0$. In conjunction with (A.5.13) we obtain

$$\int_0^t (\bar{X}^n(s))^+\,ds \leq t\Big(\frac{|x|}{\sqrt{n}} + \bar{E}^n(t)\Big) \leq t\frac{|x|}{\sqrt{n}} + \frac{t}{n}E^*(\lambda^n t) + \frac{t}{n}.$$

We have on each sample path

$$\sup_{0\leq t\leq t_0}\left|\tilde{G}^n\left(\int_0^t (\bar{X}^n(s))^+\,ds\right)\right| \leq \|\tilde{G}^n\|_{\frac{t_0|x|}{\sqrt{n}}+\frac{t_0}{n}E^*(\lambda^n t_0)+\frac{t_0}{n}}. \tag{A.5.17}$$

Since $G(\cdot)$ is a Poisson process, $\tilde{G}^n$ defined in (A.5.8) is a martingale. In addition, the random variable $E^*(\lambda^n t_0)$ is independent of $\tilde{G}^n$ and the age $A^n(0)$ associated with

the arrival process E^n. Furthermore, it follows again from Lemma 3.5 in [17] that there is some constant C independent of n, such that for $t \geq 0$,

$$\mathbb{E}\left[E^*(\lambda^n t)\right] \leq \lambda^n t + C\sqrt{\lambda^n}(1+\sqrt{t}). \tag{A.5.18}$$

Therefore we deduce from (A.5.17) that

$$\begin{aligned}
\mathbb{E}_{(a,x,z)} \left\| \tilde{G}^n \left(\int_0^t (\bar{X}^n(s))^+ \, ds \right) \right\|_{t_0}^2 &\leq \mathbb{E}_{(a,x,z)} \left[\|\tilde{G}^n\|^2_{\frac{t_0|x|}{\sqrt{n}} + \frac{t_0}{n} E^*(\lambda^n t_0) + \frac{t_0}{n}} \right] \\
&= \mathbb{E}\left[\|\tilde{G}^n\|^2_{\frac{t_0|x|}{\sqrt{n}} + \frac{t_0}{n} E^*(\lambda^n t_0) + \frac{t_0}{n}} \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\|\tilde{G}^n\|^2_{\frac{t_0|x|}{\sqrt{n}} + \frac{t_0}{n} E^*(\lambda^n t_0) + \frac{t_0}{n}} \,\middle|\, E^*(\lambda^n t_0) \right] \right] \\
&\leq \mathbb{E}\left[4\alpha \left(\frac{t_0|x|}{\sqrt{n}} + \frac{t_0}{n} E^*(\lambda^n t_0) + \frac{t_0}{n} \right) \right] \\
&\leq 4\alpha \frac{t_0|x|}{\sqrt{n}} + 4\alpha \frac{t_0}{n} \left(\lambda^n t_0 + C\sqrt{\lambda^n} + C\sqrt{\lambda^n t_0} + 1 \right),
\end{aligned}$$

where in the second last inequality we use Doob's maximal inequality for martingales (see, e.g., Revuz and Yor [89, Theorem II.1.7]), and in the last inequality we use (A.5.18). Since there is a constant C independent of n such that the sequence $\{\lambda^n/n : n \geq 1\}$ is bounded by C and

$$|x| \leq \sqrt{g(x,z)} + C \quad \text{for } (a,x,z) \in \tilde{\mathcal{S}}^n,$$

we obtain (A.5.16). On combining (A.5.14), (A.5.15) and (A.5.16), we deduce from (A.5.11) that (A.5.2) holds.

To prove (A.5.3), we start from inequality (A.5.12). Since $B^n(t)$ is the cumulative number of customers who have entered into service by time t, we obtain that

$$\bar{B}^n(t) = \frac{1}{n} B^n(t) \leq \frac{1}{n} (X^n(0))^+ + \frac{1}{n} E^n(t).$$

In addition, it is clear that $\tilde{\Phi}^{0,n}$ defined in (A.5.9) is a martingale. Thus we can proceed in a similar fashion as the proof for (A.5.16) and show that there is a constant $C(t_0)$ depending on t_0, but independent of n and any initial state (a,x,z),p such that

$$\mathbb{E}_{(a,x,z)} \|\tilde{\Phi}^{0,n}(\bar{B}^n)\|_{t_0} \leq C(t_0) + C(t_0) \frac{\sqrt[4]{g(x,z)}}{\sqrt{n}}.$$

On combining (A.5.15) we obtain (A.5.3) from (A.5.12).

APPENDIX B

APPENDIX FOR PART II

B.1 Proof of (5.2.5)

This appendix uses Theorem 5.1 to find the Laplace transform of the stationary distribution π of (Z, A) in the one-dimensional case, thereby showing in particular that Theorem 5.1 completely determines π. Writing $\mathcal{L}(\alpha, \eta)$ for the Laplace transform of π, Theorem 5.1 implies that

$$\left(\frac{1}{2}\sigma^2\alpha^2 - \alpha\theta - \eta\right)\mathcal{L}(\alpha, \eta) + \eta\mathcal{L}(0, \eta) + \alpha\theta = 0. \tag{B.1.1}$$

In particular, on setting $\eta = \frac{1}{2}\sigma^2\alpha^2 - \alpha\theta$ we get

$$\left[\frac{1}{2}\sigma^2\alpha^2 - \alpha\theta\right]\mathcal{L}\left(0, \frac{1}{2}\sigma^2\alpha^2 - \alpha\theta\right) + \alpha\theta = 0.$$

After substitution of $\alpha = (\theta + \sqrt{\theta^2 + 2\sigma^2\eta})/\sigma^2$, we find that

$$\eta\mathcal{L}(0, \eta) = -\theta\left[\frac{\theta + \sqrt{\theta^2 + 2\sigma^2\eta}}{\sigma^2}\right].$$

Substituting this back into (B.1.1) and simplifying the resulting expression, we obtain the Laplace transform given in (5.2.5).

B.2 The augmented Skorohod problem and uniqueness

In this appendix, we prove that the augmented Skorohod problem admits a unique solution. To this end, we employ a similar contraction map as in Lemma 3.6 of Mandelbaum and Ramanan [74]. Define a map Λ from $\mathbb{D}^{J\times J}$ to $\mathbb{D}^{J\times J}$ by setting, for $t \geq 0$,

$$\Lambda(\mathsf{b})_i^j(t) = \sup_{s\in\Phi_{(i)}(t)}[\chi_i^j(s) + [\tilde{\mathsf{P}}\mathsf{b}^j]_i(s)]. \tag{B.2.1}$$

Momentarily we show that Λ is a contraction map, and thus Λ has a unique fixed point b. This also implies that, defining $\mathsf{b}^{(0)} = 0$ and $\mathsf{b}^{(n)} = \Lambda(\mathsf{b}^{(n-1)})$ for $n \geq 1$, we have $\|\mathsf{b}^{(n)} - \mathsf{b}\|_T \to 0$ as $n \to \infty$ for every $T > 0$. Here and throughout this proof, we write $\|\mathsf{x}\|_T = \sup_{t\in[0,T]} |\mathsf{x}(t)|$; this should not be confused with the 1-norm and 2-norm used elsewhere in this chapter. Since χ is nonnegative and nondecreasing and $\tilde{\mathsf{P}}$ is nonnegative, we deduce that $\mathsf{b}^{(n)}$ is componentwise nonnegative and nondecreasing for each n. Therefore, we obtain that the fixed point b is also nonnegative and nondecreasing. Now let $\mathsf{a} = \chi - \tilde{\mathsf{R}}\mathsf{b}$, $\mathsf{z} = \Gamma(\mathsf{x})$, and $\mathsf{y} = \Phi(\mathsf{x})$. We next verify directly that $(\mathsf{z}, \mathsf{y}, \mathsf{a}, \mathsf{b})$ is a solution to the augmented Skorohod problem. Only the fourth and fifth requirement in Definition 5.3 are not immediate. The fourth requirement can be shown to hold using the same argument as in the proof of Lemma 5.4. For the fifth requirement, we note that if $\mathsf{z}_i(t) = 0$, (B.2.1) implies that for each j,

$$\mathsf{b}_i^j(t) = \chi_i^j(t) + (\tilde{\mathsf{P}}\mathsf{b}^j)_i(t),$$

which yields

$$\mathsf{a}_i(t) = \chi_i(t) - (\tilde{\mathsf{R}}\mathsf{b})_i(t) = \chi_i(t) + (\tilde{\mathsf{P}}\mathsf{b})_i(t) - \mathsf{b}_i(t) = 0.$$

To establish the uniqueness of solutions to the augmented Skorohod problem, we use the contraction map Λ. Suppose $(\mathsf{z}, \mathsf{y}, \mathsf{a}, \mathsf{b})$ solves the augmented Skorohod problem. Let $\tilde{\mathsf{b}} = \Lambda(\mathsf{b})$. If we can show that $\tilde{\mathsf{b}} = \mathsf{b}$, meaning b is a fixed point of Λ, then it follows from the uniqueness of the fixed point that there must be a unique solution to the augmented Skorohod problem. Suppose there exists some i, j and t_0 such that $\tilde{\mathsf{b}}_i^j(t_0) \neq \mathsf{b}_i^j(t_0)$. We discuss two cases. If $\mathsf{z}_i(t_0) = 0$, using nonnegativity and monotonicity of b, one can check from (B.2.1) that $\tilde{\mathsf{b}}_i^j(t_0) = \chi_i^j(t_0) + [\tilde{\mathsf{P}}\mathsf{b}^j]_i(t_0)$. From the definition of the augmented Skorohod problem, we also know that $\mathsf{z}_i(t_0) = 0$ implies $\mathsf{a}_i^j(t_0) = \chi_i^j(t_0) + [\tilde{\mathsf{R}}\mathsf{b}^j]_i(t_0) = 0$. Therefore, we have $\tilde{\mathsf{b}}_i^j(t_0) = \mathsf{b}_i^j(t_0)$, a contradiction. Now consider the second case where we have $\mathsf{z}_i(t_0) > 0$. If the set $\Phi_{(i)}(t_0)$ is empty, we have $\tilde{\mathsf{b}}_i^j(t_0) = \mathsf{b}_i^j(t_0) = \mathsf{b}_i^j(0) = 0$. If not, let s be the

maximal element in $\Phi_{(i)}(t_0)$. We deduce from the previous case in conjunction with the complementarily condition (5.4.1) that $\mathsf{b}_i^j(t_0) = \mathsf{b}_i^j(s) = \tilde{\mathsf{b}}_i^j(s) = \tilde{\mathsf{b}}_i^j(t_0)$. This is again a contradiction. Therefore, we obtain $\tilde{\mathsf{b}} = \mathsf{b}$ and infer that the augmented Skorohod problem has a unique solution.

It remains to show that Λ is a contraction map on $\mathbb{D}^{J\times J}$, which is equipped with the uniform norm on compact sets. As in the proof of Lemma 3.6 in Mandelbaum and Ramanan [74] we assume that, without loss of generality, the maximum row sum of $\tilde{\mathsf{P}}$ is $\eta < 1$. It is easy to verify that for any fixed $T > 0$,

$$\|\Lambda(\mathsf{b}) - \Lambda(\mathsf{b}')\|_T \leq \eta \|\mathsf{b} - \mathsf{b}'\|_T$$

for all $\mathsf{b}, \mathsf{b}' \in \mathbb{D}^{J\times J}$. Thus we have proved the existence and uniqueness of a fixed point for Λ.

APPENDIX C

APPENDIX FOR PART III

C.1 Proof of Lemma 6.1

Proof We prove (6.5.1) using a stochastic comparison approach. Equation (6.5.2) follows with a similar argument and thus is omitted here. In view of (6.2.7), it suffices to prove

$$\sup_{0\le t\le T}\left|\frac{p_A^n(t)}{n}-\frac{p_A^n(0)}{n}\right|\Rightarrow 0.$$

Given any $\epsilon>0$, it is clear that

$$\begin{aligned}&\mathbb{P}(\sup_{0\le t\le T}\left|\frac{p_A^n(t)}{n}-\frac{p_A^n(0)}{n}\right|>\epsilon)\\&\quad\le \mathbb{P}(\sup_{0\le t\le T}\tfrac{p_A^n(t)}{n}-\tfrac{p_A^n(0)}{n}>\epsilon)+\mathbb{P}(\tfrac{p_A^n(0)}{n}-\inf_{0\le t\le T}\tfrac{p_A^n(t)}{n}>\epsilon).\end{aligned}\tag{C.1.1}$$

We first show that for any small $\delta>0$, there exists N_δ such that when $n>N_\delta$,

$$\mathbb{P}(\sup_{0\le t\le T}p_A^n(t)-p_A^n(0)>n\epsilon)\le\delta.\tag{C.1.2}$$

To this end, we define an auxiliary process $\mathcal{Z}^n$ which tracks the number of sell limit orders on price levels from $p_A^n(0)$ to $p_A^n(0)+\lfloor n\epsilon\rfloor$, i.e., for each $t\ge 0$,

$$\mathcal{Z}^n(t)\stackrel{\Delta}{=}\sum_{i=p_A^n(0)}^{p_A^n(0)+\lfloor n\epsilon\rfloor}[\mathcal{X}_i^n(t)]^+.\tag{C.1.3}$$

In particular, we deduce from (6.2.6) that

$$\begin{aligned}\mathcal{Z}^n(0)&=\sum_{i=p_A^n(0)}^{p_A^n(0)+\lfloor n\epsilon\rfloor}[\mathcal{X}_i^n(0)]^+\\&=n^2\cdot\sum_{i=p_A^n(0)}^{p_A^n(0)+\lfloor n\epsilon\rfloor}\varrho(\frac{i}{n})\cdot\frac{1}{n}\\&=n^2\cdot\left(\int_p^{p+\epsilon}\varrho(x)dx+\Delta_n\right),\end{aligned}$$

where $\lim_{n\to\infty} \Delta_n = 0$. Set $\sigma^n_{\mathcal{Z}^n}$ be the first time $\mathcal{Z}^n$ reaches 0 starting from $\mathcal{Z}^n(0) > 0$, we can rewrite (C.1.2) in the following equivalent form:

$$\mathbb{P}(\sigma^n_{\mathcal{Z}^n} \leq T) \leq \delta. \tag{C.1.4}$$

To establish (C.1.4), we construct a pure-death process $\tilde{\mathcal{Z}}^n$ which is stochastically smaller than $\mathcal{Z}^n$ as defined in (C.1.3). That is, for any $t_1 < \ldots < t_j$ and any j,

$$\mathbb{E}g(\tilde{\mathcal{Z}}^n(t_1), \ldots, \tilde{\mathcal{Z}}^n(t_j)) \leq \mathbb{E}g(\mathcal{Z}^n(t_1), \ldots, \mathcal{Z}^n(t_j)),$$

for all increasing functions $g(z_1, \ldots, z_j)$. We do so by setting the birth rate of $\tilde{\mathcal{Z}}^n$ to be 0, and the death rate of $\tilde{\mathcal{Z}}^n$ to be $\Theta_{\max} \cdot i + n\Upsilon$ when the state of $\tilde{\mathcal{Z}}^n$ is i. Here

$$\Theta_{\max} = \max_{x\in[0,1]} \Theta(x).$$

In addition, we set $\tilde{\mathcal{Z}}^n(0) = \mathcal{Z}^n(0)$. Since $\mathcal{Z}^n$ increases by one at a nonnegative rate and decreases by one at a rate bounded from above by $\Theta_{\max} \cdot i + n\Upsilon$ when $\mathcal{Z}^n$ is in state i, we deduce that $\tilde{\mathcal{Z}}^n$ is stochastically smaller than $\mathcal{Z}^n$. In conjunction with the fact that first passage time is a monotone functional (see, e.g., Whitt [106]), we conclude that

$$\mathbb{P}(\sigma^n_{\mathcal{Z}^n} \leq T) \leq \mathbb{P}(\sigma^n_{\tilde{\mathcal{Z}}^n} \leq T),$$

where $\sigma^n_{\tilde{\mathcal{Z}}^n}$ is the first time $\tilde{\mathcal{Z}}^n$ reaches 0 starting from $\tilde{\mathcal{Z}}^n(0)$. Hence it suffices to prove that given any $\delta > 0$, there exists N_δ such that when $n > N_\delta$

$$\mathbb{P}(\sigma^n_{\tilde{\mathcal{Z}}^n} \leq T) \leq \delta. \tag{C.1.5}$$

We use the Lyapunov central limit theorem to prove (C.1.5). Note that given $\tilde{\mathcal{Z}}^n(0) = z$, we have

$$\sigma^n_{\tilde{\mathcal{Z}}^n} = \sum_{i=1}^{z} D_i, \tag{C.1.6}$$

where D_i is an exponential random variables with rate $i\Theta_{\max}+n\Upsilon$, and D_i represents the first passage time of $\tilde{\mathcal{Z}}^n$ from state i to $i-1$. In addition, all the D_i's are independent. Thus we have

$$\begin{aligned}
\mathbb{E}\sigma^n_{\tilde{\mathcal{Z}}^n} &= \mathbb{E}[\sum_{i=1}^{z} D_i] = \sum_{i=1}^{z} \frac{1}{i\Theta_{\max}+n\Upsilon}, \\
var(\sigma^n_{\tilde{\mathcal{Z}}^n}) &= \sum_{i=1}^{z} var(D_i) = \sum_{i=1}^{z} \frac{1}{(i\Theta_{\max}+n\Upsilon)^2}, \\
\sum_{i=1}^{z} \mathbb{E}|D_i - \mathbb{E}[D_i]|^3 &= \sum_{i=1}^{z} \frac{1}{(i\Theta_{\max}+n\Upsilon)^3} \cdot (6+\frac{6}{e}-2e).
\end{aligned}$$

One readily checks from the above equations that when $z = n^2 \cdot \left(\int_p^{p+\epsilon} \varrho(x)dx + \Delta_n\right)$ with $\lim_{n\to\infty} \Delta_n = 0$, we have some positive constants C_1, C_2, C_3 such that

$$\lim_{n\to\infty} \frac{\mathbb{E}\sigma^n_{\tilde{\mathcal{Z}}^n}}{\ln n} = C_1, \tag{C.1.7}$$

$$\lim_{n\to\infty} n \cdot var(\sigma^n_{\tilde{\mathcal{Z}}^n}) = C_2, \tag{C.1.8}$$

$$\lim_{n\to\infty} n^2 \cdot \sum_{i=1}^{z} \mathbb{E}|D_i - \mathbb{E}[D_i]|^3 = C_3. \tag{C.1.9}$$

Thus we deduce from (C.1.8) and (C.1.9) that the Lyapunov's condition holds:

$$\lim_{n\to\infty} \frac{1}{var(\sigma^n_{\tilde{\mathcal{Z}}^n})^{\frac{3}{2}}} \sum_{i=1}^{z} \mathbb{E}|D_i - \mathbb{E}[D_i]|^3 = 0.$$

It follows from the Lyapunov central limit theorem that as $n \to \infty$,

$$\frac{\sum_{i=1}^{z}[D_i - \mathbb{E}[D_i]]}{\sqrt{var(\sigma^n_{\tilde{\mathcal{Z}}})}} \Rightarrow \mathsf{D}, \tag{C.1.10}$$

where D is a standard normal random variable. Therefore we obtain from (C.1.6), (C.1.10), (C.1.7) and (C.1.8) that when $n \to \infty$,

$$\begin{aligned}
\mathbb{P}(\sigma^n_{\tilde{\mathcal{Z}}^n} \le T) &= \mathbb{P}(\sum_{i=1}^{z} D_i \le T) \\
&= \mathbb{P}\left(\frac{\sum_{i=1}^{z}[D_i - \mathbb{E}[D_i]]}{\sqrt{var(\sigma^n_{\tilde{\mathcal{Z}}^n})}} \le \frac{T - \sum_{i=1}^{z}\mathbb{E}[D_i]}{\sqrt{var(\sigma^n_{\tilde{\mathcal{Z}}^n})}}\right) \\
&\to 0.
\end{aligned}$$

This yields (C.1.5)

We next show that for $n > N_\delta$,

$$\mathbb{P}(p_A^n(0) - \inf_{0\le t\le T} p_A^n(t) > n\epsilon) \le \delta. \tag{C.1.11}$$

Note that

$$\begin{aligned} p_A^n(0) - \inf_{0\le t\le T} p_A^n(t) &\le p_A^n(0) - \inf_{0\le t\le T} p_B^n(t) \\ &= p_B^n(0) - \inf_{0\le t\le T} p_B^n(t) + p_A^n(0) - p_B^n(0). \end{aligned}$$

In view of (6.2.7), it suffices to prove

$$\mathbb{P}(p_B^n(0) - \inf_{0\le t\le T} p_B^n(t)) \le \delta.$$

This follows using a similar argument for (C.1.2).

On combining (C.1.2), (C.1.11) and (C.1.1), we have proved (6.5.1).

C.2 Proof of Lemma 6.4

We prove Lemma 6.4 in this section. We first state a sufficient condition for proving tightness of a sequence of stochastic processes living on a general Polish space. See, e.g., Kipnis and Landim [64, Section 4.1].

Lemma C.1 *Let* (E, d) *be a Polish space under metric* d*, and let* $\{\mathbb{X}_\cdot^n : n \ge 1\}$ *be a sequence of stochastic processes on* $\mathbb{D}([0,T], \mathsf{E})$*. Then* $\{\mathbb{X}_\cdot^n : n \ge 1\}$ *is tight if these two conditions hold:*

(a) For any $t \in [0,T]$ *and* $\epsilon > 0$*, there exists a compact set* $\mathsf{K}(t,\epsilon) \subset \mathsf{E}$ *such that*

$$\inf_n \mathbb{P}(\mathbb{X}_t^n \in \mathsf{K}(t,\epsilon)) > 1 - \epsilon. \tag{C.2.1}$$

(b)

$$\lim_{\sigma\to 0} \limsup_{n\to\infty} \sup_{\tau\in\Theta_T, s\le\sigma} \mathbb{P}\{\mathsf{d}(\mathbb{X}_{\tau+s}^n, \mathbb{X}_\tau^n) > \epsilon\} = 0, \tag{C.2.2}$$

for every $\epsilon > 0$*, where* Θ_T *is the family of all stopping times bounded by* T *and* $\tau + s$ *is read as* $(\tau + s) \wedge T$*.*

We also present another lemma stating for nonnegative-measure-valued process, one can understand its tightness by their actions on smooth functions. See, e.g., Kipnis and Landim [64, Section 4.1, 4.2].

Lemma C.2 *A family of nonnegative measure-valued processes* $\{\nu_\cdot^n : n \geq 1\}$ *is tight in* $\mathbb{D}([0,T], \mathcal{M}^+[0,1])$ *if* $\{\langle \nu_\cdot^n, f\rangle : n \geq 1\}$ *is tight in* $\mathbb{D}([0,T];\mathbb{R})$ *for every* $f \in C^2[0,1]$.

Proof of Lemma 6.4 Given Lemma C.2, it suffices to show, for every $f \in C^2[0,1]$, that the sequences of real-valued processes $\{\langle \zeta_\cdot^{n,+}, f\rangle : n \geq 1\}$ and $\{\langle \zeta_\cdot^{n,+}, f\rangle : n \geq 1\}$ are both tight in $\mathbb{D}([0,T];\mathbb{R})$. We concentrate below on proving the sequences of real-valued process $\{\langle \zeta_\cdot^{n,+}, f\rangle : n \geq 1\}$ is tight in $\mathbb{D}([0,T];\mathbb{R})$. The tightness of $\{\langle \zeta_\cdot^{n,-}, f\rangle : n \geq 1\}$ follows using a similar argument.

Given $f \in C^2[0,1]$, we construct a family of $C^2[0,1]$ functions f_ϵ indexed by $\epsilon > 0$ such that

$$f_\epsilon(x) = \begin{cases} f(x) & \text{if } \mathsf{p} \leq x \leq 1, \\ 0 & \text{if } 0 \leq x \leq \mathsf{p} - \epsilon, \\ \text{smooth} & \text{if } \mathsf{p} - \epsilon < x < \mathsf{p}. \end{cases}$$

Given the definition of $\zeta_\cdot^{n,+}, \zeta_\cdot^n$, and f_ϵ, one can expect the "difference" between $\langle \zeta_\cdot^{n,+}, f\rangle$ and $\langle \zeta_\cdot^n, f_\epsilon\rangle$ is small. The following lemma verifies this intuition.

Lemma C.3 *For each small* $\epsilon > 0$, *we have*

$$\begin{aligned} &\mathbb{E} \sup_{t\in[0,T]} \left|\langle \zeta_\cdot^{n,+}, f\rangle - \langle \zeta_\cdot^n, f_\epsilon\rangle\right| \\ &\quad \leq \quad o(1) + C\epsilon \cdot \left(\max_{x\in[0,1]} |\varrho(x)| + T \max_{x\in[0,1]} |\Lambda(x)| \right), \end{aligned} \tag{C.2.3}$$

where $o(1)$ *is a term going to 0 as* $n \to \infty$, *and* C *is a positive constant.*

Starting from Lemma C.3, we use Lemma C.1 and Lemma 6.3 to show the tightness of $\{\langle \zeta_\cdot^{n,+}, f\rangle : n \geq 1\}$ in $\mathbb{D}([0,T],\mathbb{R})$.

We first check condition (a) in Lemma C.1. Note that for fixed $t \in [0,T]$,

$$\begin{aligned}
&\sup_n \mathbb{P}(|\langle \zeta_t^{n,+}, f\rangle| > L) \\
&\leq \sup_n \mathbb{P}(|\langle \zeta_t^{n,+}, f\rangle - \langle \zeta_t^n, f_\epsilon\rangle| > \frac{L}{2}) + \sup_n \mathbb{P}(|\langle \zeta_t^n, f_\epsilon\rangle| > \frac{L}{2}) \\
&\leq \frac{2}{L} \sup_n \left(o(1) + C\epsilon \cdot \left(\max_{x\in[0,1]} |\varrho(x)| + T \max_{x\in[0,1]} |\Lambda(x)| \right) \right) + \sup_n \mathbb{P}(|\langle \zeta_t^n, f_\epsilon\rangle| > \frac{L}{2}),
\end{aligned}$$

where we use Markov inequality and (C.2.3) in the last inequality. Since for fixed $\epsilon > 0$, $\{\langle \zeta_t^n, f_\epsilon\rangle : n \geq 1\}$ is a tight sequence, we can pick L large to make $\sup_n \mathbb{P}(|\langle \zeta_t^{n,+}, f\rangle| > L)$ arbitrarily small.

We next check condition (b) in Lemma C.1. For the stopping times τ and $\tau + s$ bounded by T we have

$$\begin{aligned}
&\mathbb{P}\{|\langle \zeta_{\tau+s}^{n,+}, f\rangle - \langle \zeta_\tau^{n,+}, f\rangle| > c\} \\
&\leq \mathbb{P}\{|\langle \zeta_{\tau+s}^{n,+}, f\rangle - \langle \zeta_{\tau+s}^n, f_\epsilon\rangle| > \frac{c}{3}\} + \mathbb{P}\{|\langle \zeta_\tau^{n,+}, f\rangle - \langle \zeta_\tau^n, f_\epsilon\rangle| > \frac{c}{3}\} \\
&\quad + \mathbb{P}\{|\langle \zeta_{\tau+s}^n, f_\epsilon\rangle - \langle \zeta_\tau^n, f_\epsilon\rangle| > \frac{c}{3}\} \\
&\leq \frac{6}{c}\left(o(1) + C\epsilon\right) + \mathbb{P}\{|\langle \zeta_{\tau+s}^n, f_\epsilon\rangle - \langle \zeta_\tau^n, f_\epsilon\rangle| > \frac{c}{3}\}.
\end{aligned}$$

Therefore for fixed $\epsilon > 0$, we find for every $c > 0$

$$\lim_{\sigma\to 0} \limsup_{n\to\infty} \sup_{\tau\in\Theta_T, s\leq\sigma} \mathbb{P}\{|\langle \zeta_{\tau+s}^{n,+}, f\rangle - \langle \zeta_\tau^{n,+}, f\rangle| > c\} \leq \frac{6C\epsilon}{c}.$$

Since the left-hand side of the above inequality is independent of ϵ, we obtain

$$\lim_{\sigma\to 0} \limsup_{n\to\infty} \sup_{\tau\in\Theta_T, s\leq\sigma} \mathbb{P}\{|\langle \zeta_{\tau+s}^{n,+}, f\rangle - \langle \zeta_\tau^{n,+}, f\rangle| > c\} = 0,$$

after setting $\epsilon \to 0+$. The proof is therefore complete.

C.3 Proof of Lemma 6.3

Proof We use the tightness criteria as stated in Lemma C.1. The main technical part of the proof consists in showing that the generator operator of the Markov process $\mathcal{X}^n$

converges and a certain martingale vanishes when $n \to \infty$. Below we use a generic constant C which may vary from line to line, but is independent of n.

Fix a function $f \in C^2[0,1]$. For notational convenience, we write for $t \geq 0$

$$\mathcal{Y}_t^n \triangleq \langle \zeta_t^n, f \rangle \quad \triangleq \quad \frac{1}{n^2} \sum_{i=1}^{n} \mathcal{X}_i^n(t) \cdot f(\frac{i}{n}). \tag{C.3.1}$$

We first prove (C.2.1) is true for $\mathcal{Y}^n$. Since $\sup_{x \in [0,1]} |f(x)| \leq C$ for some constant C by the continuity of f, we deduce from (C.3.1) that

$$|\mathcal{Y}_t^n| \leq \frac{1}{n^2} \sum_{i=1}^{n} |\mathcal{X}_i^n(t)| \cdot |f(\frac{i}{n})| \leq \frac{C}{n^2} \sum_{i=1}^{n} |\mathcal{X}_i^n(t)|. \tag{C.3.2}$$

We now state a useful result which yields bounds on the first two moments of the (scaled) total number of limit orders in the book. The proof is given in Appendix C.9.

Lemma C.4 *Fix $T > 0$. There exists a positive constant C which depends on T but is independent of n such that*

$$\sup_n \mathbb{E}\left[\sup_{0 \leq t \leq T} \frac{1}{n^2} \sum_{i=1}^{n} |\mathcal{X}_i^n(t)| \right] \leq C, \tag{C.3.3}$$

$$\sup_n \mathbb{E}\left[\sup_{0 \leq t \leq T} \frac{1}{n^2} \sum_{i=1}^{n} |\mathcal{X}_i^n(t)| \right]^2 \leq C. \tag{C.3.4}$$

In addition, for any stopping time τ bounded by T and $s > 0$, we have

$$\sup_n \frac{1}{n^2} \mathbb{E}\left[\int_{\tau}^{\tau+s} \sum_{i=1}^{n} |\mathcal{X}_i^n(u)| du \right] \leq Cs + Cs^2. \tag{C.3.5}$$

On combining (C.3.2) with (C.3.3), we find for fixed $t \in [0,T]$

$$\sup_n \mathbb{E}|\mathcal{Y}_t^n| \leq C.$$

An application of Markov's inequality immediately yields (C.2.1).

Next we proceed to prove (C.2.2) for $\mathcal{Y}^n$. Recall from (6.7.6) that

$$\mathcal{Y}_t^n = \mathcal{Y}_0^n + \int_0^t \mathcal{L}_n F(\mathcal{X}^n(s)) ds + \mathsf{M}_t^n, \tag{C.3.6}$$

where $\mathcal{L}_n$ is the generator operator for $\mathcal{X}^n$ as given in (6.2.3), M^n is a martingale with respect to the filtration generated by $\mathcal{X}^n$, and the function F is defined in (6.7.7). Given $\epsilon > 0, s > 0$ and stopping time $\tau \leq T$, we deduce from (C.3.6) that

$$\mathbb{P}(|\mathcal{Y}^n_{\tau+s} - \mathcal{Y}^n_\tau| > \epsilon) \leq \mathbb{P}\left(\left|\int_\tau^{\tau+s} \mathcal{L}_n F(\mathcal{X}^n(u))du\right| > \epsilon/2\right) + \mathbb{P}(|\mathsf{M}^n_{\tau+s} - \mathsf{M}^n_\tau| > \epsilon/2). \tag{C.3.7}$$

Our strategy is to show for each n and stopping time τ bounded by T, there is a positive constant C independent of n such that

$$\mathbb{E}\left\{\left|\int_\tau^{\tau+s} \mathcal{L}_n F(\mathcal{X}^n(u))du\right|\right\} \leq Cs^2 + s(C + \epsilon_n), \tag{C.3.8}$$

$$\mathbb{E}\{(\mathsf{M}^n_{\tau+s} - \mathsf{M}^n_\tau)^2\} \leq \frac{1}{n^2}[Cs^2 + s(C + \hat{\epsilon}_n)], \tag{C.3.9}$$

where $\lim_{n\to 0} \epsilon_n = \lim_{n\to 0} \hat{\epsilon}_n = 0$. Applying Markov's inequality and Chebyshev's inequality, one finds

$$\lim_{\sigma\to 0} \limsup_{n\to\infty} \sup_{\tau\in\Theta_T, s\leq\sigma} \mathbb{P}(|\int_\tau^{\tau+s} \mathcal{L}_n F(\mathcal{X}^n(u))du| > \epsilon/2) = 0,$$

and

$$\lim_{\sigma\to 0} \limsup_{n\to\infty} \sup_{\tau\in\Theta_T, s\leq\sigma} \mathbb{P}(|\mathsf{M}^n_{\tau+s} - \mathsf{M}^n_\tau| > \epsilon/2) = 0.$$

On combining with (C.3.7), we obtain (C.2.2).

The rest of the proof focuses on establishing (C.3.8) and (C.3.9). We start with proving (C.3.8). For fixed n, substituting the linear function F defined in (6.7.7) into the generator (6.2.3), we obtain

$$\begin{aligned}
\mathcal{L}_n F(\mathcal{X}^n(u)) = &\sum_{k<p^n_A(u)} \left[\frac{1}{n^2} f(\frac{k}{n})\Theta^n(p^n_A(u) - k)|\mathcal{X}^n_k(u)| - \frac{1}{n^2} f(\frac{k}{n})\Lambda^n(p^n_A(u) - k)\right] \\
&+ \sum_{k>p^n_B(u)} \left[\frac{1}{n^2} f(\frac{k}{n})\Lambda^n(k - p^n_B(u)) - \frac{1}{n^2} f(\frac{k}{n})\Theta^n(k - p^n_B(u))|\mathcal{X}^n_k(u)|\right] \\
&+ \frac{1}{n^2}\left(f(\frac{p^n_B(u)}{n}) - f(\frac{p^n_A(u)}{n})\right)\Upsilon^n.
\end{aligned} \tag{C.3.10}$$

Since f and Θ are continuous functions on $[0,1]$, they are uniformly bounded by some constant C. On combining with Assumption 6.1 we deduce that

$$\begin{aligned}
|\mathcal{L}_n F(\mathcal{X}^n(u))| &\leq \sum_{k<p_A^n(u)} \left[\frac{C^2}{n^2}|\mathcal{X}_k^n(u)| + \frac{C}{n}\Lambda(\frac{p_A^n(u)-k}{n})\right] \\
&\quad + \sum_{k>p_B^n(u)} \left[\frac{C}{n}\Lambda(\frac{k-p_B^n(u)}{n}) + \frac{C^2}{n^2}|\mathcal{X}_k^n(u)|\right] + \frac{2C}{n}\Upsilon \\
&\leq \frac{C^2}{n^2}\sum_{k=1}^n |\mathcal{X}_k^n(u)| + \frac{2C}{n}\sum_{k=1}^n \Lambda(k/n) + \frac{2C}{n}\Upsilon \\
&= \frac{C^2}{n^2}\sum_{k=1}^n |\mathcal{X}_k^n(u)| + 2C\int_0^1 \Lambda(x)dx + \epsilon_n, \qquad \text{(C.3.11)}
\end{aligned}$$

where

$$\epsilon_n = \frac{2C}{n}\Upsilon + 2C\left(\frac{1}{n}\sum_{k=1}^n \Lambda(k/n) - \int_0^1 \Lambda(x)dx\right). \qquad \text{(C.3.12)}$$

Thus we conclude from (C.3.11) that

$$\begin{aligned}
\mathbb{E}\left\{\left|\int_\tau^{\tau+s} \mathcal{L}_n F(\mathcal{X}_u^n)du\right|\right\} &\leq \mathbb{E}\left[\int_\tau^{\tau+s} |\mathcal{L}_n F(\mathcal{X}_u^n)|du\right] \\
&\leq \frac{C^2}{n^2}\mathbb{E}\left[\int_\tau^{\tau+s}\sum_{k=1}^n |\mathcal{X}_k^n(u)|du\right] + s\cdot(2C\int_0^1 \Lambda(x)dx + \epsilon_n).
\end{aligned}$$

It is clear from (C.3.12) that $\lim_{n\to\infty}\epsilon_n = 0$. Hence (C.3.8) is an immediate corollary of (C.3.5) and the above inequality.

We next show (C.3.9). Using the function F in (6.7.7), we define an auxiliary process to construct martingales from the Markov chain $\mathcal{X}^n$: for each $u \geq 0$, set

$$\eta^n(u) = \mathcal{L}_n F^2(\mathcal{X}^n(u)) - 2F(\mathcal{X}^n(u))\mathcal{L}_n F(\mathcal{X}^n(u)), \qquad \text{(C.3.13)}$$

where $\mathcal{L}_n$ is the infinitesimal generator operator for $\mathcal{X}^n$ given in (6.2.3) and F is the linear function given in (6.7.7). One readily checks from (6.2.3), (6.7.7) and (C.3.13)

that

$$\eta^n(u) = \sum_{k<p_A^n(u)} \left[\frac{1}{n^4} f(\frac{k}{n})^2 \Theta^n(p_A^n(u)-k) |\mathcal{X}_k^n(u)| + \frac{1}{n^4} f(\frac{k}{n})^2 \Lambda^n(p_A^n(u)-k) \right] + \sum_{k>p_B^n(u)} \left[\frac{1}{n^4} f(\frac{k}{n})^2 \Lambda^n(k-p_B^n(u)) + \frac{1}{n^4} f(\frac{k}{n})^2 \Theta^n(k-p_B^n(u)) |\mathcal{X}_k^n(u)| \right] + \frac{\Upsilon^n}{n^4} \left(f(\frac{p_A^n(u)}{n})^2 + f(\frac{p_B^n(u)}{n})^2 \right). \tag{C.3.14}$$

Note that for each fixed n, $\{(M_t^n)^2 - \int_0^t \eta^n(u)du : t \geq 0\}$ is a martingale with respect to the filtration generated by the n-dimensional Markov chain $\mathcal{X}^n = \{\mathcal{X}^n(t) : t \geq 0\}$. See, e.g., Lemma A.1.5.1 in Kipnis and Landim [64]. This implies $\{\int_0^t \eta^n(u)du : t \geq 0\}$ is the quadratic variation process of the martingale M^n. Since τ is a stopping time, we deduce from Ito isometry that

$$\mathbb{E}\{(\mathsf{M}_{\tau+s}^n - \mathsf{M}_\tau^n)^2\} = \mathbb{E}\{\int_\tau^{\tau+s} \eta^n(u)du\}. \tag{C.3.15}$$

Using Assumption 6.1, (C.3.14) and the fact that functions f and Θ are uniformly bounded by some constant $C > 0$ on $[0,1]$, we obtain

$$\begin{aligned} \eta^n(u) &\leq \frac{C^2}{n^4} \left(C \sum_{k=1}^n |\mathcal{X}_k^n(u)| + 2 \sum_{i=1}^n \Lambda^n(i) + 2\Upsilon^n \right) \\ &= \frac{1}{n^2} \left[\frac{C^3}{n^2} \sum_{k=1}^n |\mathcal{X}_k^n(u)| + 2C^2 \cdot \frac{1}{n} \sum_{i=1}^n \Lambda(\frac{i}{n}) + \frac{2C\Upsilon}{n} \right] \\ &= \frac{1}{n^2} \left[\frac{C^3}{n^2} \sum_{k=1}^n |\mathcal{X}_k^n(u)| + 2C^2 \int_0^1 \Lambda(x)dx + \hat{\epsilon}_n \right], \end{aligned} \tag{C.3.16}$$

where

$$\hat{\epsilon}_n = \frac{2C\Upsilon}{n} + 2C^2 \cdot \left(\frac{1}{n} \sum_{i=1}^n \Lambda(\frac{i}{n}) - \int_0^1 \Lambda(x)dx \right). \tag{C.3.17}$$

It is clear that $\lim_{n\to\infty} \hat{\epsilon}_n = 0$. Therefore, we conclude from (C.3.15), (C.3.16) and (C.3.5) that (C.3.9) holds. The proof is thus complete.

C.4 Proof of Lemma 6.5

Proof It suffices to show $|\zeta_t|$ defined in (6.7.2) is absolutely continuous with respect to the Lebesgue measure. Define for $t \in [0,T]$,

$$|\zeta_t^n| \stackrel{\Delta}{=} \zeta_t^{n,+} + \zeta_t^{n,-} = \frac{1}{n^2}\sum_{i=1}^{n} |\mathcal{X}_i^n(t)| \cdot \delta_{\frac{i}{n}}. \tag{C.4.1}$$

We find from (6.3.4) that for $f \in \mathcal{C}[0,1]$,

$$|\langle |\zeta_t^n|, f\rangle| \Rightarrow |\langle |\zeta_t|, f\rangle| \quad \text{as } n \to \infty.$$

Since $|\zeta_t|$ is deterministic, we obtain when $n \to \infty$

$$|\langle |\zeta_t^n|, f\rangle| \to |\langle |\zeta_t|, f\rangle| \quad \text{in probability.} \tag{C.4.2}$$

Now using (C.4.1) we obtain

$$\begin{aligned} |\langle |\zeta_t^n|, f\rangle| &= \left|\frac{1}{n^2}\sum_{i=1}^{n} |\mathcal{X}_i^n(t)| \cdot f(\frac{i}{n})\right| \\ &\leq \left|\frac{1}{n^2}\sum_{i=1}^{n} |\mathcal{X}_i^n(0)| \cdot f(\frac{i}{n})\right| + \left|\frac{1}{n^2}\sum_{i=1}^{n} |\mathcal{E}_i^n(t)| \cdot f(\frac{i}{n})\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} |\varrho(\frac{i}{n})| \cdot |f(\frac{i}{n})| + \frac{1}{n^2}\sum_{i=1}^{n} |\mathcal{E}_i^n(t)| \cdot \left|f(\frac{i}{n})\right|, \end{aligned} \tag{C.4.3}$$

where $\mathcal{E}_i^n(t)$ is the total number of new limit orders submitted at price level i. For fixed n, one readily checks that $\mathcal{E}_i^n$ is stochastically bounded by a Poisson process with rate $n\Lambda_{\max}$ where

$$\Lambda_{\max} \stackrel{\Delta}{=} \max_{x\in[0,1]} \Lambda(x).$$

Thus we can bound the first moment of the last display in (C.4.3) by

$$\frac{1}{n}\sum_{i=1}^{n} |\varrho(\frac{i}{n})| \cdot |f(\frac{i}{n})| + \Lambda_{\max} t \cdot \frac{1}{n}\sum_{i=1}^{n} |f(\frac{i}{n})|. \tag{C.4.4}$$

Since the above two terms are both integrable, we deduce that the sequence $\{|\langle |\zeta_t^n|, f\rangle| : n \geq 1\}$ is uniformly integrable. On combining (C.4.2), (C.4.3) and (C.4.4) we conclude that

$$|\langle |\zeta_t|, f\rangle| \leq \int_0^1 |\varrho(x)| \cdot |f(x)| dx + \Lambda_{\max} t \cdot \int_0^1 |f(x)| dx.$$

Therefore the measures $|\zeta_t|$ and ζ_t are absolutely continuous with respect to the Lebesgue measure. The proof is thus complete.

C.5 Proof of Lemma 6.8

Proof Given any $\epsilon > 0$, since for fixed $n \geq 1$, $\{\mathsf{M}_t^n : t \geq 0\}$ is a martingale with respect to the natural filtration generated by $\mathcal{X}^n = \{\mathcal{X}_1^n, \ldots, \mathcal{X}_n^n\}$, we deduce that

$$\begin{aligned}
\mathbb{P}(\sup_{0\leq t\leq T} |\mathsf{M}_t^n| > \epsilon) &\leq \frac{1}{\epsilon^2}\mathbb{E}\sup_{0\leq t\leq T} |\mathsf{M}_t^n|^2 \\
&\leq \frac{4}{\epsilon^2}\sup_{0\leq t\leq T}\mathbb{E}|\mathsf{M}_t^n|^2 \\
&= \frac{4}{\epsilon^2}\mathbb{E}|\mathsf{M}_T^n|^2 \\
&\leq \frac{4}{n^2\epsilon^2}[CT^2 + T(C+\hat{\epsilon}_n)],
\end{aligned}$$

where C is a positive constant and $\hat{\epsilon}_n = 0$ is given in (C.3.17). Here, the first inequality is a result of Chebyshev's inequality, the second inequality follows from Doob's maximal inequality for martingales, the next equality comes from the fact that $|\mathsf{M}_t^n|^2$ is submartingale, and the last inequality is obtained from (C.3.9). Therefore we find (6.7.8).

C.6 Proof of Lemma 6.9

Proof The key idea is to use Lemma 6.1, which shows the scaled best bid and ask prices converge, and allows us to decouple the dynamics of buy and sell sides of order book in the $n \to \infty$ limit. Given the generator in (C.3.10), we first show

$$\sup_{0\leq s\leq T}\left|\sum_{k>p_B^n(s)}\left[\frac{1}{n^2}f(\frac{k}{n})\Lambda^n(k-p_B^n(s))\right] - \int_{\mathsf{p}}^1 f(x)\Lambda(x-\mathsf{p})dx\right| \Rightarrow 0 \quad \text{as } n\to\infty. \tag{C.6.1}$$

By Assumption 6.1 it is equivalent to show as $n \to \infty$

$$\sup_{0\leq s\leq T}\left|\sum_{k>p_B^n(s)}\left[\frac{1}{n}f\left(\frac{k}{n}\right)\Lambda\left(\frac{k}{n}-\frac{p_B^n(s)}{n}\right)\right] - \int_{\mathsf{p}}^1 f(x)\Lambda(x-\mathsf{p})dx\right| \Rightarrow 0. \tag{C.6.2}$$

To this end, we define the following functions for $z \in [0,1]$

$$\begin{aligned} h_n(z) &= \sum_{\frac{k}{n}>z} \left[\frac{1}{n} f\left(\frac{k}{n}\right) \Lambda\left(\frac{k}{n} - z\right) \right], \\ h(z) &= \int_z^1 f(x)\Lambda(x-z)dx. \end{aligned}$$

It is clear that

$$\lim_{n\to\infty} h_n(z) = h(z) \quad \text{for each } z \in [0,1].$$

Since Λ and f are both continuous functions, it follows that h are continuous and thus uniformly continuous on the compact interval $[0,1]$. In addition, from the fact that $\Lambda(0) = 0$, one also readily checks that h_n are continuous and thus uniformly continuous on the compact interval $[0,1]$ for each n. This implies

$$\lim_{n\to\infty} h_n(z_n) = h(z) \quad \text{if } \lim_{n\to\infty} z_n = z \in [0,1].$$

In conjunction with Lemma 6.1, we thus obtain from the generalized continuous mapping theorem (see, e.g., Whitt [107, Theorem 3.4.4]) that (C.6.2) holds.

Using a similar argument we find that

$$\sup_{0\le s\le T} \left| \sum_{k<p_A^n(s)} \left[\frac{1}{n^2} f(\frac{k}{n}) \Lambda^n(p_A^n(s) - k) \right] - \int_0^{\mathsf{p}} f(x)\Lambda(\mathsf{p} - x)dx \right| \Rightarrow 0 \quad \text{as } n \to \infty. \tag{C.6.3}$$

We next prove when $n \to \infty$,

$$\sup_{0\le s\le T} \left| \sum_{k>p_B^n(s)} \left[\frac{1}{n^2} f\left(\frac{k}{n}\right) \Theta^n(k - p_B^n(s)) \cdot \mathcal{X}_k^n(s) \right] - \int_{\mathsf{p}}^1 f(x)\Theta(x - \mathsf{p})d\zeta_s(x) \right| \Rightarrow 0. \tag{C.6.4}$$

By Assumption 6.1 and (6.3.4) it suffices to show as $n \to \infty$

$$\sup_{0\le s\le T} \left| \sum_{k>p_B^n(s)} \left[\frac{1}{n^2} f\left(\frac{k}{n}\right) \Theta\left(\frac{k - p_B^n(s)}{n}\right) \cdot \mathcal{X}_k^n(s) \right] - \int_{\mathsf{p}}^1 f(x)\Theta(x - \mathsf{p})d\zeta_s^n(x) \right| \Rightarrow 0.$$

Using the definition of ζ^n, it is equivalent to show

$$\sup_{0\le s\le T}\left|\sum_{\frac{k}{n}>\frac{P_B^n(s)}{n}}\left[\frac{1}{n^2}f\left(\frac{k}{n}\right)\Theta\left(\frac{k}{n}-\frac{p_B^n(s)}{n}\right)\cdot\mathcal{X}_k^n(s)\right]\right.$$
$$\left.-\sum_{\frac{k}{n}>\mathsf{p}}\left[\frac{1}{n^2}f\left(\frac{k}{n}\right)\Theta\left(\frac{k}{n}-\mathsf{p}\right)\cdot\mathcal{X}_k^n(s)\right]\right|\Rightarrow 0. \quad \text{(C.6.5)}$$

To prove the above equation, we define

$$\hat{\Theta}(x)=\begin{cases}\Theta(x) & \text{if } x\in[0,1],\\ 0 & \text{if } x\in[-1,0).\end{cases}$$

The function $\hat{\Theta}$ is continuous and thus uniformly continuous on $[-1,1]$ since Θ is continuous on $[0,1]$ and $\Theta(0)=0$. Now we can rewrite (C.6.5) as follows:

$$\sup_{0\le s\le T}\left|\sum_{k=1}^{n}\left\{\frac{1}{n^2}f\left(\frac{k}{n}\right)\cdot\left[\hat{\Theta}\left(\frac{k}{n}-\frac{p_B^n(s)}{n}\right)-\hat{\Theta}\left(\frac{k}{n}-\mathsf{p}\right)\right]\cdot\mathcal{X}_k^n(s)\right\}\right|\Rightarrow 0. \quad \text{(C.6.6)}$$

Indeed we establish below when $n\to\infty$,

$$\mathbb{E}\left[\sup_{0\le s\le T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\hat{\Theta}\left(\frac{k}{n}-\frac{p_B^n(s)}{n}\right)-\hat{\Theta}\left(\frac{k}{n}-\mathsf{p}\right)\right|\cdot|\mathcal{X}_k^n(s)|\right]\to 0, \quad \text{(C.6.7)}$$

from which (C.6.6) follows by application of Markov's inequality and using the fact that f is bounded on $[0,1]$. Given any $\epsilon>0$, to show (C.6.7), we split it into two parts:

$$\mathbb{E}\left[\sup_{0\le s\le T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\hat{\Theta}\left(\frac{k}{n}-\frac{p_B^n(s)}{n}\right)-\hat{\Theta}\left(\frac{k}{n}-\mathsf{p}\right)\right|\cdot|\mathcal{X}_k^n(s)|\,;\,\sup_{0\le s\le T}\left|\frac{p_B^n(s)}{n}-\mathsf{p}\right|<\delta\right], \quad \text{(C.6.8)}$$

and

$$\mathbb{E}\left[\sup_{0\le s\le T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\hat{\Theta}\left(\frac{k}{n}-\frac{P_B^n(s)}{n}\right)-\hat{\Theta}\left(\frac{k}{n}-\mathsf{p}\right)\right|\cdot|\mathcal{X}_k^n(s)|\,;\,\sup_{0\le s\le T}\left|\frac{P_B^n(s)}{n}-\mathsf{p}\right|\ge\delta\right], \quad \text{(C.6.9)}$$

where $\delta>0$ is a small constant depending on ϵ and $\hat{\Theta}$. By the uniform continuity of the function $\hat{\Theta}$, we can choose δ such that (C.6.8) is upper bounded by

$$\epsilon\cdot\frac{1}{n^2}\mathbb{E}\left[\sup_{0\le s\le T}\sum_{k=1}^{n}|\mathcal{X}_k^n(s)|\right],$$

which is further bounded by ϵC uniformly for all n after invoking Lemma C.4. To bound (C.6.9), we note that $\hat{\Theta}$ is uniformly bounded by some constant C on $[-1,1]$. This implies (C.6.9) is upper bounded by

$$2C \cdot \mathbb{E}\left[\sup_{0\le s\le T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\mathcal{X}_k^n(s)\right|; \sup_{0\le s\le T}\left|\frac{P_B^n(s)}{n}-\mathsf{p}\right|\ge\delta\right]. \tag{C.6.10}$$

Now (C.3.4) in Lemma C.4 implies the sequence $\{\sup_{0\le s\le T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\mathcal{X}_k^n(s)\right| : n\ge 1\}$ is uniformly bounded in L^2, thus $\{\sup_{0\le s\le T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\mathcal{X}_k^n(s)\right| : n\ge 1\}$ is uniformly integrable. On combining with Lemma 6.1 we deduce that (C.6.10) is bounded by $2\epsilon C$. Therefore, we have proved (C.6.7), and (C.6.4) follows.

Similarly we can show as $n\to\infty$,

$$\sup_{0\le s\le T}\left|\sum_{k<P_A^n(s)}\left[\frac{1}{n^2}f\left(\frac{k}{n}\right)\Theta^n(p_A^n(s)-x)\cdot\mathcal{X}_k^n(s)\right]-\int_0^{\mathsf{p}}f(x)\Theta(\mathsf{p}-x)d\zeta_s(x)\right|\Rightarrow 0. \tag{C.6.11}$$

Finally, using Assumption 6.1 and the fact that f is bounded by C on $[0,1]$, we obtain

$$\sup_{0\le s\le T}\left|\frac{\Upsilon^n}{n^2}\left(f(\frac{p_B^n(s)}{n})-f(\frac{p_A^n(s)}{n})\right)\right|\le\frac{\Upsilon^n\cdot 2C}{n^2}\Rightarrow 0\quad\text{as } n\to\infty. \tag{C.6.12}$$

On combining (C.6.1), (C.6.3), (C.6.4), (C.6.11), and (C.6.12), we obtain (6.7.9).

C.7 Proof of Lemma 6.10

Proof Given Lemma 6.5, we have $\zeta_t^+(\mathsf{p})=\zeta_t^-(\mathsf{p})=0$. To show (ζ_t^+,ζ_t^-) is the Hahn-Jordan decomposition of the signed measure ζ_t, it suffices to show that the support of ζ_t^+ is contained in $[\mathsf{p},1]$ and the support of ζ_t^- is contained in $[0,\mathsf{p}]$. We first prove ζ_t^+ concentrates on $[\mathsf{p},1]$. Suppose this is not true, then there exists some $x\in[0,p)$ such that for every open neighbourhood $\mathcal{O}_x\subset[0,p)$ of x, we have $\zeta_t^+(\mathcal{O}_x)>0$. By the weak convergence (6.7.1), we obtain

$$\zeta_t^{n_k,+}(\mathcal{O}_x)\Rightarrow\zeta_t^+(\mathcal{O}_x)\quad\text{as } n_k\to\infty.$$

However, from Lemma 6.1 and the definition of $\zeta_t^{n_k,+}$, one deduces that $\zeta_t^{n_k,+}(\mathcal{O}_x) \Rightarrow 0$ as $n_k \to \infty$. This is a contradiction. We therefore conclude that the support of ζ_t^+ is contained in $[\mathsf{p}, 1]$. The proof for ζ_t^- is similar and thus is omitted.

C.8 Proof of Lemma C.3

Proof We use Lemma 6.1 and the stochastic comparison result for the limit order arrival processes on each price level. Note that

$$\begin{aligned}
&\mathbb{E} \sup_{t\in[0,T]} \left|\langle \zeta_\cdot^{n,+}, f\rangle - \langle \zeta_\cdot^n, f_\epsilon\rangle\right| \\
&= \frac{1}{n^2}\mathbb{E} \sup_{t\in[0,T]} \left| \sum_{i>p_B^n(t)} \mathcal{X}_i^n(t) f(\frac{i}{n}) - \sum_{i=1}^{n} \mathcal{X}_i^n(t) f_\epsilon(\frac{i}{n}) \right| \\
&= \frac{1}{n^2}\mathbb{E} \sup_{t\in[0,T]} \left| \sum_{i>p_B^n(t)} \mathcal{X}_i^n(t) f(\frac{i}{n}) - \sum_{i\geq np} \mathcal{X}_i^n(t) f(\frac{i}{n}) - \sum_{i\in(n\mathsf{p}-n\epsilon, n\mathsf{p})} \mathcal{X}_i^n(t) f_\epsilon(\frac{i}{n}) \right| \\
&\leq \frac{1}{n^2}\mathbb{E} \sup_{t\in[0,T]} \left| \sum_{i>p_B^n(t)} \mathcal{X}_i^n(t) f(\frac{i}{n}) - \sum_{i\geq np} \mathcal{X}_i^n(t) f(\frac{i}{n}) \right| + \frac{1}{n^2}\mathbb{E} \sup_{t\in[0,T]} \left| \sum_{i\in(n\mathsf{p}-n\epsilon, n\mathsf{p})} \mathcal{X}_i^n(t) f_\epsilon(\frac{i}{n}) \right|.
\end{aligned} \tag{C.8.1}$$

Since f_ϵ can be constructed such that this family is uniformly bounded by some constant C on $[0, 1]$, we are able to bound the second term in (C.8.1) by

$$\frac{C}{n^2} \cdot \mathbb{E} \sup_{t\in[0,T]} \sum_{i\in(n\mathsf{p}-n\epsilon, n\mathsf{p})} |\mathcal{X}_i^n(t)|,$$

which is further bounded by

$$\frac{C}{n^2} \cdot \mathbb{E} \sum_{i\in(n\mathsf{p}-n\epsilon, n\mathsf{p})} |\mathcal{X}_i^n(0)| + \frac{C}{n^2} \cdot \mathbb{E} \sum_{i\in(n\mathsf{p}-n\epsilon, n\mathsf{p})} |\mathcal{E}_i^n(T)|.$$

By Assumption 6.2 and the stochastic comparison result for $\mathcal{E}_i^n$, we deduce the above display is bounded by

$$C\epsilon \cdot \left(\max_{x\in[0,1]} |\varrho(x)| + T \max_{x\in[0,1]} |\Lambda(x)| \right).$$

We next bound the first term in (C.8.1). We split it into two parts and study them separately. Fixing some $\delta > 0$, we first note that

$$\frac{1}{n^2}\mathbb{E}\left[\sup_{t\in[0,T]}\left|\sum_{i>p_B^n(t)}\mathcal{X}_i^n(t)f(\frac{i}{n})-\sum_{i\geq n\mathsf{p}}\mathcal{X}_i^n(t)f(\frac{i}{n})\right| : \sup_{t\in[0,T]}\left|\frac{p_B^n(t)}{n}-\mathsf{p}\right|>\delta\right]$$
$$\leq 2C\cdot\mathbb{E}\left[\sup_{0\leq s\leq T}\sum_{i=1}^{n}\frac{1}{n^2}\left|\mathcal{X}_i^n(s)\right| ; \sup_{t\in[0,T]}\left|\frac{p_B^n(t)}{n}-\mathsf{p}\right|\geq\delta\right]. \quad \text{(C.8.2)}$$

Now (C.3.4) in Lemma C.4 implies the sequence $\{\sup_{0\leq s\leq T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\mathcal{X}_k^n(s)\right| : n\geq 1\}$ is uniformly bounded in L^2, thus $\{\sup_{0\leq s\leq T}\sum_{k=1}^{n}\frac{1}{n^2}\left|\mathcal{X}_k^n(s)\right| : n\geq 1\}$ is uniformly integrable. On combining with Lemma 6.1 we deduce that for any $\delta>0$, the display (C.8.2) goes to 0 when $n\to\infty$. We next bound

$$\frac{1}{n^2}\mathbb{E}\left[\sup_{t\in[0,T]}\left|\sum_{i>p_B^n(t)}\mathcal{X}_i^n(t)f(\frac{i}{n})-\sum_{i\geq n\mathsf{p}}\mathcal{X}_i^n(t)f(\frac{i}{n})\right| : \sup_{t\in[0,T]}\left|\frac{p_B^n(t)}{n}-\mathsf{p}\right|<\delta\right].$$

Using a similar argument as before, this term is upper bounded by

$$C\delta\cdot\left(\max_{x\in[0,1]}|\varrho(x)|+T\max_{x\in[0,1]}|\Lambda(x)|\right).$$

Therefore we find

$$\mathbb{E}\sup_{t\in[0,T]}\left|\langle\zeta_\cdot^{n,+},f\rangle-\langle\zeta_\cdot^{n},f_\epsilon\rangle\right|$$
$$\leq\ o(1)+C(\epsilon+\delta)\cdot\left(\max_{x\in[0,1]}|\varrho(x)|+T\max_{x\in[0,1]}|\Lambda(x)|\right),$$

where $o(1)$ represents the term in (C.8.2), which goes to 0 as $n\to\infty$. Moreover, since the left-hand side of the above inequality does not depend on δ, we can let $\delta\to 0+$ and obtain (C.2.3) for each small $\epsilon>0$. We thus have completed the proof.

C.9 Proof of Lemma C.4

Proof The key idea of the proof is to stochastically bound the limit order arrival processes on each price level. We first show (C.3.4). It is clear that (C.3.4) implies (C.3.3). Set

$$\mathcal{N}^n(t)=\sum_{i=1}^{n}|\mathcal{X}_i^n(t)|.$$

$\mathcal{N}^n(t)$ is the total number of limit orders on the order book at time t. Fox fixed n, let $\mathcal{E}^n(\cdot)$ be the aggregated arrival process to the order book, i.e., $\mathcal{E}^n(t)$ is the total number of limit order arrivals on all price levels during $[0,t]$. One can construct a Poisson process $\tilde{\mathcal{E}}^n(\cdot)$ with rate $r_n \stackrel{\Delta}{=} 2\sum_{i=1}^n \Lambda^n(i)$ on the same probability space as $\mathcal{N}^n(\cdot)$ such that

$$\tilde{\mathcal{E}}^n(t,\omega) \geq \mathcal{E}^n(t,\omega) \quad \text{for every sample path } \omega. \tag{C.9.1}$$

Now set

$$\tilde{\mathcal{N}}^n(t) = \mathcal{N}^n(0) + \tilde{\mathcal{E}}^n(t); \tag{C.9.2}$$

then $\tilde{\mathcal{N}}^n$ is a pure birth process and we find from (C.9.1) that

$$\tilde{\mathcal{N}}^n \geq \mathcal{N}^n \quad \text{for every sample path.} \tag{C.9.3}$$

In particular, for fixed $T > 0$

$$\sup_{0\leq t\leq T} \mathcal{N}^n(t) \leq \sup_{0\leq t\leq T} \tilde{\mathcal{N}}^n(t) = \tilde{\mathcal{N}}^n(0) + \tilde{\mathcal{E}}^n(T) \quad \text{for every sample path.}$$

This implies

$$\mathbb{E}[\sup_{0\leq t\leq T} \mathcal{N}^n(t)^2] \leq \mathbb{E}[\mathcal{N}^n(0) + \tilde{\mathcal{E}}^n(T)]^2 \leq 2\mathbb{E}[\mathcal{N}^n(0)]^2 + 2\mathbb{E}[\tilde{\mathcal{E}}^n(T)]^2. \tag{C.9.4}$$

We first use Assumption 6.2 to bound $\mathbb{E}[\mathcal{N}^n(0)]^2$. Since for fixed n, $|\mathcal{X}_i^n(0)| = n|\varrho(i/n)|$, we deduce

$$\mathbb{E}[\mathcal{N}^n(0)]^2 = \mathbb{E}\left(\sum_{i=1}^n |\mathcal{X}_i^n(0)|\right)^2 = m_n^2, \tag{C.9.5}$$

where

$$m_n = n\sum_{i=1}^n |\varrho(i/n)|. \tag{C.9.6}$$

One readily checks that there exists some $C > 0$ such that

$$\sup_n \frac{m_n}{n^2} = \sup_n \frac{1}{n}\sum_{i=1}^n |\varrho(i/n)| \leq \int_0^1 |\varrho(u)|du + C. \tag{C.9.7}$$

In addition, since φ is continuous, it is integrable. Hence we deduce from the above inequality and (C.9.5) that there exists some constant $C > 0$ such that

$$\sup_n \frac{1}{n^4}\mathbb{E}[\mathcal{N}^n(0)]^2 \leq C. \tag{C.9.8}$$

To bound the second term on the right-hand side of (C.9.4), we note $\tilde{\mathcal{E}}^n$ is a Poisson process with rate r_n. Thus we have

$$\mathbb{E}[\tilde{\mathcal{E}}^n(T)]^2 = r_n T + r_n^2 T^2. \tag{C.9.9}$$

Using Assumption 6.1 we find there exists some constant $C > 0$ such that for each n

$$r_n = 2\sum_{i=1}^n \Lambda^n(i) = 2n^2 \sum_{i=1}^n \frac{1}{n}\Lambda(\frac{i}{n}) \leq 2n^2(\int_0^1 \Lambda(u)du + C). \tag{C.9.10}$$

On combining with (C.9.9) we have

$$\sup_n \frac{1}{n^4}\mathbb{E}[\tilde{\mathcal{E}}^n(T)]^2 \leq C + CT^2. \tag{C.9.11}$$

Now (C.3.4) follows from (C.9.4), (C.9.8) and (C.9.11).

We next prove (C.3.5). It follows from (C.9.2) and (C.9.3) that

$$\frac{1}{n^2}\mathbb{E}\left[\int_\tau^{\tau+s} \mathcal{N}^n(u)du\right] \quad \leq \quad s \cdot \frac{1}{n^2}\mathbb{E}\left[\mathcal{N}^n(0)\right] + \frac{1}{n^2}\mathbb{E}\left[\int_\tau^{\tau+s} \tilde{\mathcal{E}}^n(u)du\right]. \tag{C.9.12}$$

The first term on the right-hand side of (C.9.12) is bounded by Cs due to (C.9.7) and

$$\mathbb{E}[\mathcal{N}^n(0)] = \mathbb{E}\left(\sum_{i=1}^n |\mathcal{X}_i^n(0)|\right) = m_n, \tag{C.9.13}$$

where m_n is given in (C.9.6). For the second term, we have for each n,

$$\begin{aligned}
\frac{1}{n^2}\mathbb{E}\left[\int_\tau^{\tau+s} \tilde{\mathcal{E}}^n(u)du\right] &= \mathbb{E}\left[\mathbb{E}[\int_\tau^{\tau+s} \frac{1}{n^2}\tilde{\mathcal{E}}^n(u)du|\tau]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\int_\tau^{\tau+s} \frac{r_n u}{n^2}du\right)\right] \\
&= \frac{r_n}{2n^2}\mathbb{E}[(\tau+s)^2 - \tau^2] \\
&= \frac{r_n}{2n^2}(s^2 + 2s\mathbb{E}\tau),
\end{aligned} \tag{C.9.14}$$

where the second equality follows from Fubini's theorem and the fact that $\tilde{\mathcal{E}}^n$ is a Poisson process with rate r_n. Noting that τ is bounded by T, therefore we deduce from (C.9.14) and (C.9.10) that the second term on the right-hand side of (C.9.12) is bounded by $C(s + s^2)$ for some constant C. On combining (C.9.13) and (C.9.6), we have obtained (C.3.5). The proof of the lemma is complete.

REFERENCES

[1] Abergel, F. and Jedidi, A., "A mathematical approach to order book modelling," in *Econophysics of Order-driven Markets*, pp. 93–107, Springer, 2011.

[2] Abramowitz, M. and Stegun, I. A., eds., *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover Publications Inc., 1992. Reprint of the 1972 edition.

[3] Aksin, Z., Armony, M., and Mehrotra, V., "The modern call center: A multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, no. 6, pp. 665–688, 2007.

[4] Alfonsi, A. and Acevedo, J. I., "Optimal execution and price manipulations in time-varying limit order books," *arXiv preprint arXiv:1204.2736*, 2012.

[5] Alfonsi, A., Fruth, A., and Schied, A., "Optimal execution strategies in limit order books with general shape functions," *Quantitative Finance*, vol. 10, no. 2, pp. 143–157, 2010.

[6] Anderson, R. F. and Orey, S., "Small random perturbation of dynamical systems with reflecting boundary," *Nagoya Math. J.*, vol. 60, pp. 189–216, 1976.

[7] Asmussen, S. and Glynn, P. W., *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.

[8] Bak, P., Paczuski, M., and Shubik, M., "Price variations in a stock market with many agents," *Physica A: Statistical Mechanics and its Applications*, vol. 246, no. 3, pp. 430–453, 1997.

[9] Berman, A. and Plemmons, R. J., *Nonnegative matrices in the mathematical sciences*, vol. 9 of *Classics in Applied Mathematics*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1994. Revised reprint of the 1979 original.

[10] Biais, B., Hillion, P., and Spatt, C., "An empirical analysis of the limit order book and the order flow in the paris bourse," *Journal of Finance*, vol. 50, no. 5, pp. 1655–1689, 2012.

[11] Billingsley, P., *Convergence of Probability Measures*. New York: John Wiley & Sons Inc., second ed., 1999.

[12] Bouchaud, J.-P., Mézard, M., and Potters, M., "Statistical properties of stock order books: empirical results and models," *Quantitative Finance*, vol. 2, no. 4, pp. 251–256, 2002.

[13] Bovier, A. and Cerny, J., "Hydrodynamic limit for the $A + B \to \emptyset$ model," *Preprint*, 2006.

[14] Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V., *Linear Matrix Inequalities in System and Control Theory*, vol. 15 of *SIAM Studies in Applied Mathematics*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1994.

[15] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L., "Statistical analysis of a telephone call center," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 36–50, 2005.

[16] Budhiraja, A. and Lee, C., "Stationary distribution convergence for generalized Jackson networks in heavy traffic," *Mathematics of Operations Research*, vol. 34, no. 1, pp. 45–56, 2009.

[17] Budhiraja, A. and Ghosh, A. P., "Diffusion approximations for controlled stochastic networks: An asymptotic bound for the value function," *The Annals of Applied Probability*, vol. 16, no. 4, pp. 1962–2006, 2006.

[18] Chakraborti, A., Muni Toke, I., Patriarca, M., and Abergel, F., "Econophysics review: I. Empirical facts," *Quant. Finance*, vol. 11, no. 7, 2011.

[19] Chakraborti, A., Muni Toke, I., Patriarca, M., and Abergel, F., "Econophysics review: II. Agent-based models," *Quant. Finance*, vol. 11, no. 7, pp. 1013–1041, 2011.

[20] Cont, R. and De Larrard, A., "Order book dynamics in liquid markets: limit theorems and diffusion approximations," *Available at SSRN 1757861*, 2011.

[21] Cont, R., Stoikov, S., and Talreja, R., "A stochastic model for order book dynamics," *Oper. Res.*, vol. 58, no. 3, pp. 549–563, 2010.

[22] Dai, J. G., "On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models," *Annals of Applied Probability*, vol. 5, pp. 49–77, 1995.

[23] Dai, J. G. and Dieker, A. B., "Nonnegativity of solutions to the basic adjoint relationship for some diffusion processes," *Queueing System*, vol. 68, pp. 295–303, 2011.

[24] Dai, J. G., He, S., and Tezcan, T., "Many-server diffusion limits for $G/Ph/n + GI$ queues," *Annals of Applied Probability*, vol. 20, no. 5, pp. 1854–1890, 2010.

[25] Dai, J. G. and He, S., "Many-server queues with customer abandonment: Numerical analysis of their diffusion models," *Stochastic Systems*, 2013. To appear.

[26] Dai, J. G. and Meyn, S. P., "Stability and convergence of moments for multiclass queueing networks via fluid limit models," *IEEE Transactions on Automatic Control*, vol. 40, pp. 1889–1904, 1995.

[27] Dai, J. G. and Prabhakar, B., "The throughput of data switches with and without speedup," *IEEE INFOCOM*, pp. 556–564, 2000.

[28] Davis, M. H. A., "Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models," *Journal of Royal Statist. Soc. series B*, vol. 46, pp. 353–388, 1984.

[29] Dawson, D., *Measure-valued Markov processes*. Springer, 1993.

[30] Dębicki, K., Dieker, A. B., and Rolski, T., "Quasi-product forms for Lévy-driven fluid networks," *Mathematics of Operations Research*, vol. 32, p-p. 629–647, 2007.

[31] Del Barrio, E. and van de Geer, S. A., *Lectures on empirical processes: theory and statistical applications*. European Mathematical Society, 2007.

[32] Dieker, A. B., Ghosh, S., and Squillante, M. S., "Optimal resource capacity management for stochastic networks," *Preprint*, 2013.

[33] Dieker, A. B. and Shin, J., "From local to global stability in stochastic processing networks through quadratic Lyapunov functions," *Preprint*, 2012.

[34] Duistermaat, J. J. and Kolk, J. A. C., *Distributions*. Boston, MA: Birkhäuser, 2010.

[35] Dupuis, P. and Ishii, H., "On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications," *Stochastics*, vol. 35, pp. 31–62, 1991.

[36] Dupuis, P. and Williams, R. J., "Lyapunov functions for semimartingale reflecting Brownian motions," *Ann. Probab.*, vol. 22, no. 2, pp. 680–702, 1994.

[37] Foucault, T., Kadan, O., and Kandel, E., "Limit order book as a market for liquidity," *Review of Financial Studies*, vol. 18, no. 4, pp. 1171–1217, 2005.

[38] Fruth, A., Schöneborn, T., and Urusov, M., "Optimal trade execution and price manipulation in order books with time-varying liquidity," *Mathematical Finance*, 2013.

[39] Gamarnik, D. and Goldberg, D., "Steady-state $GI/GI/N$ queue in the Halfin-Whitt regime," *Annals of Applied Probability*, 2013. To appear.

[40] Gamarnik, D. and Momčilović, P., "Steady-state analysis of a multi-server queue in the Halfin-Whitt regime," *Advances in Applied Probability*, vol. 40, pp. 548–577, 2008.

[41] Gamarnik, D. and Stolyar, A. L., "Multiclass multiserver queueing system in the Halfin-Whitt heavy traffic regime: asymptotics of the stationary distribution," *Queueing Systems*, vol. 71, no. 1-2, pp. 25–51, 2012.

[42] Gamarnik, D. and Zeevi, A., "Validity of heavy traffic steady-state approximation in generalized Jackson networks," *Annals of Appllied Probability*, vol. 16, no. 1, pp. 56–90, 2006.

[43] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers: tutorial, review, and research prospects," *Manufacturing and Service Operations Management*, vol. 5, pp. 79–141, 2003.

[44] Glasserman, P., *Gradient Estimation Via Perturbation Analysis*. Springer, 1991.

[45] Glasserman, P., "Regenerative derivatives of regenerative sequences," *Adv. in Appl. Probab.*, vol. 25, no. 1, pp. 116–139, 1993.

[46] Glasserman, P., "Perturbation analysis of production networks," in *Stochastic modeling and analysis of manufacturing systems* (Yao, D. D., ed.), pp. 233–280, Springer, New York, 1994.

[47] Gould, M., Porter, M., Williams, S., McDonald, M., Fenn, D., and Howison, S., "Limit order books," *Available at SSRN 1970185*, 2012.

[48] Gurvich, I., "Diffusion models and steady-state approximations for exponentially ergodic markovian queues," *Preprint*, 2013.

[49] Gurvich, I., "Validity of heavy-traffic steady-state approximations in multiclass queueing networks: The case of queue-ratio disciplines," *Mathematics of Operations Research*, 2013. to appear.

[50] Halfin, S. and Whitt, W., "Heavy-traffic limits for queues with many exponential servers," *Operations Research*, vol. 29, pp. 567–588, 1981.

[51] Harrison, J. M. and Williams, R. J., "Multidimensional reflected Brownian motions having exponential stationary distributions," *Ann. Probab.*, vol. 15, pp. 115–137, 1987.

[52] Harrison, J. M. and Williams, R. J., "Brownian models of open queueing networks with homogeneous customer populations," *Stochastics*, vol. 22, pp. 77–115, 1987.

[53] Heidergott, B., *Max-plus linear stochastic systems and perturbation analysis*. New York: Springer, 2006.

[54] Horn, R. A. and Johnson, C. R., *Topics in matrix analysis*. Cambridge, UK: Cambridge University Press, 1994.

[55] Horsty, U. and Paulsen, M., "A law of large numbers for limit order books," *Preprint*, 2013.

[56] Jacod, J. and Shiryaev, A. N., *Limit theorems for stochastic processes*. Berlin: Springer, second ed., 2003.

[57] Kalman, R. E., "Lyapunov functions for the problem of Lur′e in automatic control," *Proc. Nat. Acad. Sci. U.S.A.*, vol. 49, pp. 201–205, 1963.

[58] Kang, W. and Ramanan, K., "Characterization of stationary distributions of reflected diffusions," *arXiv preprint arXiv:1204.4969*, 2012.

[59] Karatzas, I. and Shreve, S. E., *Brownian motion and stochastic calculus*. New York: Springer, second ed., 1991.

[60] Katsuda, T., "State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic," *Queueing Systems*, vol. 65, no. 3, pp. 237–273, 2010.

[61] Kella, O. and Whitt, W., "Stability and structural properties of stochastic storage networks," *J. Appl. Probab.*, vol. 33, no. 4, pp. 1169–1180, 1996.

[62] Khasminskii, R. Z., *Stochastic stability of differential equations*, vol. 66 of *Stochastic Modelling and Applied Probability*. New York: Springer-Verlag, 2011. second edition.

[63] King, C. and Nathanson, M., "On the existence of a common quadratic Lyapunov function for a rank one difference," *Linear Algebra Appl.*, vol. 419, no. 2-3, pp. 400–416, 2006.

[64] Kipnis, C. and Landim, C., *Scaling limits of interacting particle systems*, vol. 320. Springer, 1999.

[65] Kleinrock, L., *Communication Nets; Stochastic Message Flow and Delay*. McGraw-Hill Book Company, New York, 1964.

[66] Konstantopoulos, T. and Last, G., "On the use of Lyapunov function methods in renewal theory," *Stochastic Process. Appl.*, vol. 79, no. 1, pp. 165–178, 1999.

[67] Kotelenez, P., "Stochastic flows and signed measure valued stochastic partial differential equations," *Th. Stoch. Process*, 2010.

[68] Kruk, Ł., "Functional limit theorems for a simple auction," *Mathematics of Operations Research*, vol. 28, no. 4, pp. 716–751, 2003.

[69] Kruk, Ł., "Limiting distribution for a simple model of order book dynamics," *Central European Journal of Mathematics*, vol. 10, no. 6, pp. 2283–2295, 2012.

[70] Lasry, J.-M. and Lions, P.-L., "Mean field games," *Japanese Journal of Mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

[71] Lehalle, C.-A., Guéant, O., and Razafinimanana, J., "High-frequency simulations of an order book: a two-scale approach," in *Econophysics of Order-driven Markets*, pp. 73–92, Springer, 2011.

[72] Liggett, T. M., *Interacting particle systems*. Springer, 2005.

[73] Mandelbaum, A. and Massey, W. A., "Strong approximations for time-dependent queues," *Math. Oper. Res.*, vol. 20, no. 1, pp. 33–64, 1995.

[74] Mandelbaum, A. and Ramanan, K., "Directional derivatives of oblique reflection maps," *Mathematics of Operations Research*, vol. 35, no. 3, pp. 527–558, 2010.

[75] Meyn, S. P. and Down, D., "Stability of generalized Jackson networks," *Annals of Applied Probability*, vol. 4, pp. 124–148, 1994.

[76] Meyn, S. and Tweedie, R. L., *Markov chains and stochastic stability*. Cambridge: Cambridge University Press, second ed., 2009.

[77] Meyn, S. P. and Tweedie, R. L., "Stability of Markovian processes III: Foster-Lyapunov criteria for continuous time processes," *Adv. Appl. Probab.*, vol. 25, pp. 518–548, 1993.

[78] Neuts, M. F., *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Dover publications, 1995.

[79] Obizhaeva, A. A. and Wang, J., "Optimal trading strategy and supply/demand dynamics," *Journal of Financial Markets*, 2012.

[80] Pang, G. and Yao, D. D., "Heavy-traffic limits for a multiclass many-server queueing network with switchovers," *Advances in Applied Probability*, 2013. To appear.

[81] Parlour, C. and Seppi, D., "Limit order markets: A survey," *Handbook of Financial Intermediation and Banking*, vol. 5, 2008.

[82] Peszat, S. and Zabczyk, J., "Strong Feller property and irreducibility for diffusions on Hilbert spaces," *Ann. Probab.*, vol. 23, no. 1, pp. 157–172, 1995.

[83] Pollett, P. K., "Optimal capacity assignment in general queueing networks," in *Optimization: structure and applications*, vol. 32 of *Springer Optim. Appl.*, pp. 261–272, New York: Springer, 2009.

[84] Potters, M. and Bouchaud, J., "More statistical properties of order books and price impact," *Physica A: Statistical Mechanics and its Applications*, vol. 324, no. 1, pp. 133–140, 2003.

[85] Predoiu, S., Shaikhet, G., and Shreve, S., "Optimal execution in a general one-sided limit-order book," *SIAM Journal on Financial Mathematics*, vol. 2, pp. 183–212, 2011.

[86] Protter, P. E., *Stochastic integration and differential equations*, vol. 21 of *Stochastic Modelling and Applied Probability*. Berlin: Springer-Verlag, 2005. Second edition. Version 2.1, Corrected third printing.

[87] Puhalskii, A. A. and Reiman, M. I., "The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime," *Advances in Applied Probability*, vol. 32, pp. 564–595, 2000. Correction: **36**, 971 (2004).

[88] Ramanan, K., "Reflected diffusions defined via the extended Skorokhod map," *Electron. J. Probab.*, vol. 11, pp. 934–992, 2006.

[89] Revuz, D. and Yor, M., *Continuous martingales and Brownian motion*. Springer Verlag, 1999.

[90] Rogers, L. C. G. and Williams, D., *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library, Cambridge: Cambridge University Press, 2000. Itô calculus, Reprint of the second (1994) edition.

[91] Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J., *Stochastic processes for insurance and finance*. Wiley, 2009.

[92] Roşu, I., "A dynamic model of the limit order book," *Review of Financial Studies*, vol. 22, no. 11, pp. 4601–4641, 2009.

[93] Rudin, W., *Functional analysis*. International Series in Pure and Applied Mathematics, New York: McGraw-Hill Inc., second ed., 1991.

[94] Russell, J. and Kim, T., "A new model for limit order book dynamics," *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, pp. 354–364, 2010.

[95] Shorten, R., Corless, M., Wulff, K., Klinge, S., and Middleton, R., "Quadratic stability and singular SISO switching systems," *IEEE Trans. Automat. Control*, vol. 54, no. 11, pp. 2714–2718, 2009.

[96] Shorten, R., Wirth, F., Mason, O., Wulff, K., and King, C., "Stability criteria for switched and hybrid systems," *SIAM Rev.*, vol. 49, no. 4, pp. 545–592, 2007.

[97] Shorten, R. N. and Narendra, K. S., "On common quadratic Lyapunov functions for pairs of stable LTI systems whose system matrices are in companion form," *IEEE Trans. Automat. Control*, vol. 48, no. 4, pp. 618–621, 2003.

[98] Smith, E., Farmer, J., Gillemot, L., and Krishnamurthy, S., "Statistical theory of the continuous double auction," *Quantitative finance*, vol. 3, no. 6, pp. 481–514, 2003.

[99] Stanley, R. P., *Enumerative combinatorics. Vol. 1.* Cambridge: Cambridge University Press, 1997.

[100] Stolyar, A. L., "On the stability of multiclass queueing networks: a relaxed sufficient condition via limiting fluid processes," *Markov Processes and Related Fields*, vol. 1, pp. 491–512, 1995.

[101] Tassiulas, L. and Ephremides, A., "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, no. 12, pp. 1936–1948, 1992.

[102] Tezcan, T., "Optimal control of distributed parallel server systems under the Halfin and Whitt regime," *Mathematics of Operations Research*, vol. 33, no. 1, pp. 51–90, 2008.

[103] Toth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J., and Bouchaud, J.-P., "Anomalous price impact and the critical nature of liquidity in financial markets," *Physical Review X*, vol. 1, no. 2, p. 021006, 2011.

[104] Varadarajan, V., "Weak convergence of measures on separable metric spaces," *Sankhyā: The Indian Journal of Statistics (1933-1960)*, vol. 19, no. 1/2, pp. 15–22, 1958.

[105] Wein, L. M., "Capacity allocation in generalized Jackson networks," *Oper. Res. Lett.*, vol. 8, no. 3, pp. 143–146, 1989.

[106] Whitt, W., "Comparing counting processes and queues," *Adv. in Appl. Probab.*, vol. 13, no. 1, pp. 207–220, 1981.

[107] Whitt, W., *Stochastic-process limits.* New York: Springer, 2002.

[108] Williams, R. J., "Reflected Brownian motion with skew symmetric data in a polyhedral domain," *Probab. Theory Related Fields*, vol. 75, pp. 459–485, 1987.

[109] Ye, H.-Q. and Yao, D. D., "Diffusion limit of a two-class network: stationary distributions and interchange of limits," *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 2, pp. 18–20, 2010.

[110] Yin, G. G. and Zhu, C., *Hybrid switching diffusions.* New York: Springer, 2010.