MANY-SERVER QUEUES WITH CUSTOMER ABANDONMENT

A Thesis
Presented to
The Academic Faculty

by

Shuangchi He

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

Georgia Institute of Technology
August 2011

MANY-SERVER QUEUES WITH CUSTOMER ABANDONMENT

Approved by:

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

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

Professor Hayriye Ayhan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Tolga Tezcan
Simon Graduate School of Business
University of Rochester

Professor Robert D. Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 30 June 2010

ACKNOWLEDGEMENTS

I have had a phenomenal time during my four years in Georgia Tech. I attribute my good fortune to these wonderful people.

First and foremost, I would like to thank my advisor, Professor Jim Dai, for his guidance and support. I will always remember him as "the advisor who made everything possible". His friendliness, wisdom, spontaneity, and attitude of hopefulness made him unique. I have learned a lot and will benefit for life from both his instructions and his personality.

Many thanks go to my committee members, Professors Hayriye Ayhan, Bob Foley, Anton Kleywegt, and Tolga Tezcan. They have provided me with invaluable guidance and friendliness during my studies and contributed much to my thesis work.

I would also like to thank Professor Olav Kallenberg, who opened the door to the world of probability for me. His guidance made me decide to explore a new field.

These acknowledgements would be far from complete without thanks to Professors Ton Dieker and Bert Zwart, for their sage advice during my study.

Finally, I would like to express my gratitude to my parents and my wife, Fan, whose love and support is the source of my strength.

My studies were funded by National Science Foundation under Grants CMMI 0727400 and 1030589.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Customer call centers with hundreds of agents working in parallel are ubiquitous in many industries. These systems have a large amount of daily traffic that is stochastic in nature. It becomes more and more difficult to manage a call center because of its increasingly large scale and the stochastic variability in arrival and service processes. In call center operations, customer abandonment is a key factor and may significantly impact the system performance. It must be modeled explicitly in order for an operational model to be relevant for decision making.

In this thesis, a large-scale call center is modeled as a queue with many parallel servers. To model the customer abandonment, each customer is assigned a patience time. When his waiting time for service exceeds his patience time, a customer abandons the system without service. We develop analytical and numerical tools for analyzing such a queue.

We first study a sequence of $G/G/n + GI$ queues, where the customer patience times are independent and identically distributed (iid) following a general distribution. The focus is the abandonment and the queue length processes. We prove that under certain conditions, a deterministic relationship holds asymptotically in diffusion scaling between these two stochastic processes, as the number of servers goes to infinity.

Next, we restrict the service time distribution to be a phase-type distribution with d phases. Using the aforementioned asymptotic relationship, we prove limit theorems for $G/Ph/n+GI$ queues in the quality- and efficiency-driven (QED) regime. In particular, the limit process for the customer number in each phase is a d-dimensional piecewise Ornstein–Uhlenbeck (OU) process.

Motivated by the diffusion limit process, we propose two approximate models for a $GI/Ph/n+GI$ queue. In each model, a d-dimensional diffusion process is used to approximate the dynamics of the queue. These two models differ in how the patience time distribution is built into them. The first diffusion model uses the patience time density at zero and the second one uses the entire patience time distribution. We also develop a numerical algorithm to analyze these diffusion models. The algorithm solves the stationary distribution of each model. The computed stationary distribution is used to estimate the queue's performance. A crucial part of this algorithm is to choose an appropriate reference density that controls the convergence of the algorithm. We develop a systematic approach to constructing a reference density. With the proposed reference density, the algorithm is shown to converge quickly in numerical experiments. These experiments also show that the diffusion models are good approximations of queues with a moderate to large number of servers.

CHAPTER I

INTRODUCTION

Customer call centers have become an important part of the service economy in a modern society. To take advantage of the economy of scale, call centers with hundreds of agents are ubiquitous in many industries. These systems face a large amount of daily traffic that is intrinsically stochastic and has temporal variations. In a call center, a customer waiting for service may hang up the phone before being served. This is called *customer abandonment*. Such a phenomenon is common because customers usually have limited patience. Customer expectation demands that a proper staffing level be maintained in the call center so that most customers are served without waiting for a long time and only a small fraction of customers abandon the system. As pointed by [18], customer abandonment is a crucial factor for call center operations. It may significantly impact the system performance and must be modeled explicitly in order for an operational model to be relevant for decision making. In this thesis, we develop analytical and numerical approaches to analyzing the mathematical models of a large-scale call center.

1.1 Overview

A queue with many parallel servers has been used extensively to model a customer call center. See, e.g., [1, 17] for surveys. In this thesis, we model customer abandonment by assigning each customer a random patience time. When a customer's waiting time exceeds his patience time, the customer abandons the system without any service. The mathematical model of a multi-server queue with customer abandonment is detailed in Chapter 2. As the queue is usually invisible to customers in a call center, it is reasonable to assume that the patience times are independent and

identically distributed (iid).

The exact analysis of a many-server queue has been mostly restricted to the $M/M/n+M$ model (also called the Erlang-A model in the literature) that has a Poisson arrival process and exponential service and patience time distributions [18]. As pointed out in [6], however, the service time distribution of a call center appears to follow a log-normal distribution. In [61], the patience time distribution of a call center has also been observed to be far from exponential. With a general service or patience time distribution, there is no finite-dimensional Markovian representation of the queue. Except computer simulation, there are no methods that are able to analyze such a queue either analytically or numerically. To deal with this challenge, the following strategies are adopted in this thesis for analyzing a many-server queue.

First, we focus on the queues operated in the *quality- and efficiency-driven (QED) regime*: Such a queue has a large number of parallel servers and a high arrival rate; the arrival rate and the service capacity is approximately balanced so that the server utilization is close to one. As argued by [18], such a system is characterized by short customer waiting times and a small fraction of abandonment, even though the server utilization is high. Hence, both quality and efficiency can be achieved in this queue. To mathematically define the QED regime, we consider a sequence of queues indexed by the number of servers n. For the nth system, its arrival rate λ^n depends on n. The arrival rate $\lambda^n \to \infty$ as $n \to \infty$, whereas the service time and the patience time distributions do not change with n. We use $1/\mu$ to denote the mean service time of each customer, and define the traffic intensity of the nth system as $\rho^n = \lambda^n/(n\mu)$. We assume that

$$\lim_{n\to\infty} \sqrt{n}\,(1-\rho^n) = \beta \quad \text{for some } \beta \in \mathbb{R}. \tag{1.1}$$

When condition (1.1) holds, the *sequence of queues* is said to be in the QED regime. In Chapter 3, we establish an asymptotic relationship that characterizes the abandonment processes of a sequence of $G/G/n+GI$ queues in the QED regime. More

specifically, we prove in Theorem 3.1 that

$$\frac{1}{\sqrt{n}} \sup_{0\le t\le T} \left| A^n(t) - \alpha \int_0^t Q^n(s)\,\mathrm{d}s \right| \to 0 \quad \text{in probability as } n \to \infty, \tag{1.2}$$

where $A^n(t)$ is the number of customers who have abandoned the nth queue by time t, $Q^n(t)$ is the number of waiting customers of the nth queue at time t, and α is the density at zero of the patience time distribution. For this relationship to hold, a key assumption is the stochastic boundedness of the diffusion-scaled queue length processes. To allow one to verify this assumption, we establish a comparison result in Theorem 3.2. This theorem states that at any time, the queue length in a queue with customer abandonment is always dominated by the queue length in a queue with longer service times and no abandonment. Using (1.2), one can replace the abandonment process by the queue length integral with respect to time in the system equation. The error resulting from this replacement is asymptotically negligible in diffusion scaling. This step is the key to proving the many-server heavy traffic limit theorems in Chapter 4. Such an approximation is also used in a diffusion model for many-server queues in Chapter 5.

Next, we restrict the service time distribution to be phase-type. Since phase-type distributions can approximate any positive-valued distribution, such a queue is still relevant to practical systems. The resulting system is a $G/Ph/n+GI$ queue. Four limit theorems are proved in Chapter 4 for a sequence of $G/Ph/n+GI$ queues in the QED regime. Theorem 4.1 states that after proper centering and scaling, the customer-count processes and the server-allocation processes converge jointly to a continuous multidimensional Markov process in distribution. Then, we prove a state space collapse result in Theorem 4.2: At any time, if the waiting customers are classified according to their first service phases, then they are distributed among these phases following the initial distribution of the phase-type distribution. Using the above two theorems, we derive a diffusion limit in Theorem 4.3 for the number of

customers in each phase. This limit process is a multidimensional piecewise Ornstein–Uhlenbeck (OU) process. This diffusion limit motivates us to explore approximate models in Chapter 5. The relationship between the virtual waiting time processes and the queue length processes is set up in Theorem 4.4.

The limit process in Theorem 4.3 implies that in the QED regime, the dynamics of a many-server queue can be approximated by a diffusion process. In Chapter 5, we propose two approximate models for a $GI/Ph/n+GI$ queue. In each model, a multidimensional diffusion process is used to represent the scaled customer numbers among service phases. The diffusion models are obtained by replacing certain scaled renewal processes by Brownian motions. The difference between the two diffusion models lies in how the patience time distribution is built into them. The first diffusion model uses the patience time density at zero and the second one uses the entire patience time distribution. Since the exact analysis of a many-server queue is difficult, we choose to analyze the diffusion models to obtain performance estimates for the queue. To this end, the stationary distribution of a diffusion model needs to be computed. Except for special cases, the stationary distribution of a diffusion process has no explicit formula. We develop a numerical algorithm for solving the stationary distributions of the diffusion models. The computed stationary distribution is used to estimate the performance measures of the many-server queue. The proposed algorithm follows the framework set up in [10]. The starting point of the algorithm is the basic adjoint relationship that characterizes the stationary distribution of a diffusion process. A crucial part of the proposed algorithm is to choose an appropriate reference density that controls the convergence of the algorithm. The reference density is required to have a comparable or slower decay rate than the (unknown) stationary density. To facilitate the selection of such a reference density, we make a conjecture on the tail behavior of the limit queue length process for many-server queues with customer abandonment. We conjecture that the limit queue length process has a

Gaussian tail that depends on the service time distribution only through its first two moments. With this tail, we construct a product-form reference density that make the algorithm converge quickly. Numerical experiments show that the diffusion models are good approximations for queues with a moderate to large number of servers.

1.2 Contributions and related work

Zeltyn and Mandelbaum studied a large number of data sets from call centers in [61]. They observed a linear relationship between the abandonment rate and the mean waiting time. The relationship (1.2), which is proved in Theorem 3.1 of this thesis, can be used to justify this observation in the QED regime: It follows from (1.2) that

$$A^n(t) \approx \alpha \int_0^t Q^n(s)\,\mathrm{d}s, \tag{1.3}$$

where $\int_0^t Q^n(s)\,\mathrm{d}s$ can be interpreted as the cumulative waiting time of all customers. If both sides of (1.3) are divided by the number of customer arrivals by time t, the approximation turns out to be

$$\text{the abandonment fraction} \approx \alpha \times \text{the mean waiting time.}$$

Because the abandonment rate is the product of the abandonment fraction and the arrival rate, then

$$\text{the abandonment rate} \approx \lambda^n \alpha \times \text{the mean waiting time.}$$

In [61], the authors also studied steady-state quantities for $M/M/n+GI$ queues in the QED regime. In their results, the patience time distribution affects the limiting quantities only through its density at zero.

A limit theorem for a sequence of $G/GI/n+GI$ queues in the QED regime was proved in [37] by Mandelbaum and Momčilović. Although Chapter 3 of this thesis and the work of [37] are contemporary, independent studies, there is a significant overlap between the results of them. For example, Corollary 3 of [37] gives a relationship

between the abandonment processes and the queue length processes, which is similar to (1.2). Their Proposition 1, similar to our Theorem 3.2, gives a comparison between queues with and without abandonment. Also, Corollary 1 in their work is similar to Proposition 3.3 of this thesis.

The two studies, however, differ significantly, both philosophically and in terms of assumptions and proof techniques. We believe that our results have laid a framework for a modular approach to proving many-server limit theorems for queues with customer abandonment in the QED regime: (a) first, prove a limit theorem for queues without customer abandonment using a continuous-mapping approach; (b) then, use the asymptotic relationship (1.2) and a modified map to prove the corresponding limit theorem for queues with customer abandonment. This modular approach is carried out in Chapter 4 for proving limit theorems for a sequence of $G/Ph/n+GI$ queues. Further, we believe that the limit theorem of [37] for one-dimensional customer-count processes can be proved in a simpler approach using our Theorem 3.1 and the limit theorem in [47] which is for a sequence of $G/GI/n$ queues in the QED regime. Indeed, using the limit theorem of [47] as well as our comparison result in Theorem 3.2, we can readily see that the stochastic boundedness assumption is satisfied for the $G/GI/n+GI$ queues in the QED regime. Recently, a measure-valued heavy-traffic limit has been reported in [28] by Kaspi and Ramanan for a sequence of $G/GI/n$ queues. It is expected that our Theorem 3.1 can be used to generalize their result to the $G/GI/n+GI$ model. Besides these philosophical differences, our Theorem 3.1 differs from Corollary 3 of [37] in the following aspects. First, their corollary is a weak convergence result, whereas our asymptotic relationship (1.2) is a stronger result at a sample path level. Second, their corollary assumes iid service times, whereas we assume nothing on service times as long as the stochastic boundedness assumption holds. Third, we impose much weaker assumptions on arrival processes. They assume that each arrival process in the sequence has a time homogeneous arrival rate and

the sequence of arrival processes satisfies a certain functional central limit theorem (FCLT). In contrast, we assume (3.6) and (3.7) for the arrival processes that allow for non-homogeneous arrival rates and batch arrivals. These two features often exclude a functional central limit for the arrival processes. Please refer to Section 3.1 for more details. Of course, we need the stochastic boundedness of the diffusion-scaled queue length processes. This assumption implicitly requires that the sequence of queues be not overloaded in the limit.

A key insight of this thesis, as well as in [37, 61], is that in the QED regime, the exact patience time distribution is irrelevant as long as customer abandonment is explicitly built into the model. This phenomenon is in sharp contrast to the one found in [58] when the systems are operated in an overloaded regime known as the *efficiency-driven (ED) regime*. The system performance there depends crucially upon the patience time distribution and a fluid model is shown to be able to capture that dependency. In particular, it was demonstrated in [3] by Bassamboo and Randhawa that for $M/M/n+GI$ queues with certain performance measures and patience time distributions, the optimized staffing levels surprisingly drive the queues to the ED regime. In such a case, a fluid model provides accurate approximations for the performance measures and the approximation error does not increase with the system size n.

The asymptotic relationship (1.2) is also the key to proving the heavy-traffic limit theorems for a sequence of $G/Ph/n+GI$ queues in Chapter 4. These limit theorems extend the results in [43] that were proved by Puhalskii and Reiman for $G/Ph/n$ queues without abandonment. These theorems also extend the work of [18] and [57] that was established for queues with exponential service time distributions to phase-type service time distributions.

In addition to these limit theorems, the techniques used in the proofs are innovative. A sample-path representation is established in Section 4.4 for $G/Ph/n+GI$

queues. The sample-path argument has been explored previously for strong approximations in the setting of Markovian networks in [35] and in the setting of general state-dependent networks in [36]. The representation derived in Section 4.4 allows us to obtain the customer-count and the server-allocation processes as a map of primitive processes with a random time change. These primitive processes are either assumed or proved to satisfy FCLTs. In our continuous-mapping approach, we have heavily exploited some maps from $\mathbb{D}^{d+1}$ to $\mathbb{D}^{d+1}$, where d is the number of phases in the service time distribution. Variants of these maps have been employed in [12, 35, 41, 47, 54], among others. We use these maps not just in diffusion scaling but also in fluid scaling. Using a map twice, one for each scaling, allows us to obtain diffusion limits as a simple consequence of the standard continuous-mapping theorem and the random-time-change theorem (see, e.g., [4, 14] for these theorems). More specifically, by using the continuous-mapping approach, we first prove a heavy-traffic limit for $G/Ph/n$ queues without customer abandonment. As a consequence, the stochastic boundedness assumption on the diffusion-scaled queue length processes holds for $G/Ph/n$ queues, and thus holds for $G/Ph/n+GI$ queues by the comparison result in Theorem 3.2. This allows us to exploit the asymptotic relationship (1.2) to replace the abandonment process by an integral of the queue length process. Then, we can prove the corresponding heavy traffic limit for $G/Ph/n + GI$ queues with abandonment by applying the continuous-mapping approach again. A conventional heavy traffic limit theorem was proved in the seminal paper [48] by Reiman for generalized Jackson networks. Our approach resembles the work of [25], which also uses a multi-dimensional Skorohod map twice and provides a significant simplification of Reiman's original proof.

The previous studies [18, 20, 56] all use Stone's theorem to prove diffusion limit theorems. Stone's theorem is set up for convergence of Markov chains to a diffusion process [52]. This setting makes the generalization to non-renewal arrival processes

difficult. The continuous-mapping approach was also used in [43] by Puhalskii and Reiman for the $GI/Ph/n$ model. They employed a different sample-path representation for the customer-count process. Their representation requires them to extensively use martingale FCLTs in their proofs, whereas our approach uses the standard FCLT for random walks and Poisson processes. A number of sample-path representations and martingale proofs are reviewed in [41] for many-server heavy traffic limits, and the proof techniques for establishing martingale FCLTs are surveyed in [59]. Our proofs show that for the queues with a phase-type service time distribution, there is a general approach to proving limit theorems in the QED regime, without employing martingale FCLTs.

The diffusion limits for many-server queues can be traced back to [20] by Halfin and Whitt, where a diffusion limit was established in the QED regime for $GI/M/n$ queues. Their results were generalized by Puhalskii and Reiman in [43] for the $GI/Ph/n$ model. A diffusion limit was proved in [18] for the $M/M/n+M$ model, allowing for customer abandonment. This result was generalized by Whitt in [57] to the $G/M/n+M$ model. In the same paper, Whitt proved a limit process for the $G/H_2^*/n$ model; this limit is not a diffusion process but a simple transformation of it is a diffusion process. The diffusion limit for $G/Ph/n+GI$ queues was proved in [11]; this result is presented in Chapter 4 of the current thesis. Recently, a diffusion limit for $GI/M/n+GI$ queues was proved by Reed and Tezcan in [45]. In their framework, a refined limit process is obtained by scaling the patience time hazard rate function. For the overloaded $M/M/n+M$ model, it was demonstrated in [56] by Whitt that a certain fluid approximation can be useful in predicting the steady-state performance of the many-server system. Whitt further demonstrated that a diffusion limit provides a refined approximation. A diffusion limit for overloaded $G/Ph/n+M$ queues was set up in [11], which generalizes the results in [56].

For the more general $G/GI/n$ model, a many-server limit process was proved by

Reed in [47] for the customer-count process in the QED regime. His assumption on the service time distribution is completely general and the limit process is not a Markov process. This work was generalized in [37] that allows for customer abandonment. For the overloaded $G/GI/n$ model, a finite-dimensional-distribution limit was proved in [44] for the customer-count process. A limit theorem was proved in [24] for the $GI/D/n$ model with deterministic service times. In [16], Gamarnik and Momčilović studied the steady-state distribution of the limit process of the $GI/GI/n$ model, where the service times are lattice-valued on a finite support. When the service time distribution is general, measure-valued processes have been used to give a Markovian description of the queue. A measure-valued fluid limit was obtained in [29] for the $G/GI/n$ model, and a similar limit was obtained in [26] for the $G/GI/n+GI$ model with customer abandonment. In [62], Zhang derived a similar measure-valued fluid limit independently. Their work partially justifies the fluid model in [58].

The diffusion limits proved in Theorem 4.3 and in [45] motivate the two approximate models presented in Chapter 5. The major contributions there are the diffusion models and the proposed reference densities that are crucial to the numerical algorithm for computing the stationary distributions of the diffusion models. Similar diffusion models can be traced back to [21] by Harrison and Nguyen for multiclass open queueing networks. Their diffusion models are semimartingale reflected Brownian motions (SRBMs) and are rooted in the conventional heavy traffic limit theorems that are pioneered in [23] for serial networks and in [48] for single-class networks. See [60] for a survey of limit theorems in the literature. For a two-dimensional SRBM living in a rectangle, an algorithm was proposed in [9] for computing its stationary distribution. In [10], the algorithm was extended for an SRBM living in an orthant. To deal with the unbounded state space, the notion of a reference density was first introduced there. The authors of [10] used global polynomials to approximate the stationary density. With this choice, the algorithm sometimes appears numerically

unstable. In such a case, the round-off error may dominate the approximation error while the approximation error is still significant. In [50], Shen et al. extended [9] to a hypercube state space of an arbitrary dimension. They used a finite element method to avoid numerical instability. Their algorithm sometimes converges slowly because they did not explore a reference density. A linear programming algorithm for computing the stationary distribution of a diffusion process was proposed in [49] by Saure, Glynn, and Zeevi. Both SRBMs in an orthant and a diffusion approximation of many-server queues with two priority classes were investigated in their paper. Like the role of a reference density, it appears that the re-scaling of variables is essential to the convergence of their algorithm.

1.3 Organization

The rest of this thesis is organized as follows. In Chapter 2, we give the mathematical definition of a multi-server queue and present several basic results. Chapter 3 focuses on the $G/G/n+GI$ model and is dedicated to proving the asymptotic relationship (1.2), which characterizes the abandonment processes in the QED regime. We prove limit theorems in Chapter 4 for a sequence of $G/Ph/n+GI$ queues in the QED regime. Two diffusion models for $GI/Ph/n+GI$ queues are proposed in Chapter 5; in the same chapter, we develop a numerical algorithm for analyzing the diffusion models. The thesis concludes in Chapter 6 and future directions are suggested.

1.4 Notation

All random variables and stochastic processes are defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ unless otherwise specified. In this probability space, $\mathbb{E}[\cdot]$ is reserved for expectation and 1_χ is the indicator function on Ω of a set $\chi \in \mathscr{F}$, i.e., $1_\chi(\omega) = 1$ if $\omega \in \chi$ and $1_\chi(\omega) = 0$ if $\omega \notin \chi$.

The symbols $\mathbb{Z}$, $\mathbb{Z}_+$, $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{R}_+$ are used to denote the sets of integers, nonnegative integers, positive integers, real numbers, and nonnegative real numbers, respectively. For $d, m \in \mathbb{N}$, $\mathbb{R}^d$ denotes the d-dimensional Euclidean space and $\mathbb{R}^{d\times m}$ denotes the space of $d \times m$ real matrices. We use $C_b^2(\mathbb{R}^d)$ to denote the set of real-valued functions on $\mathbb{R}^d$ that are twice continuously differentiable with bounded first and second derivatives. The space of functions $f : \mathbb{R}_+ \to \mathbb{R}^d$ that are right-continuous on $[0, \infty)$ and have left limits on $(0, \infty)$ is denoted by $\mathbb{D}^d$.

For $f \in \mathbb{D}^d$, $f(t-)$ denotes its left limit at $t > 0$. Given another $\check{f} \in \mathbb{D}$ that is nondecreasing and takes values in $\mathbb{R}_+$, $f \circ \check{f}$ denotes the composed function in $\mathbb{D}$ with $(f \circ \check{f})(t) = f(\check{f}(t))$ for $t \geq 0$. For a sequence of random elements $\{X^n : n \in \mathbb{N}\}$ 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 whose sample paths are in $\mathbb{D}^d$ is considered to be a $\mathbb{D}^d$-valued random element. The space $\mathbb{D}^d$ is endowed with the Skorohod J_1-topology (see [4, 14]). For an index set J and a set of random variables $\{Y_j : j \in J\}$, $\sigma\{Y_j : j \in J\}$ is the σ-field generated by $\{Y_j : j \in J\}$.

All vectors are envisioned as column vectors. For a d-dimensional vector $v \in \mathbb{R}^d$, we use v_j to denote its jth entry and $\mathrm{diag}(v)$ for the $d \times d$ diagonal matrix with jth diagonal entry v_j. For a matrix M, M' denotes its transpose, M_{jk} denotes its (j, k)th entry, and $|M| = \max\{|M_{jk}| : j, k = 1, \ldots, d|\}$. We reserve I for the $d \times d$ identity matrix, e for the d-dimensional vector of ones, and e^j for the d-dimensional vector with jth entry one and all other entries zero.

Given $z, w \in \mathbb{R}$, $z \vee w = \max\{z, w\}$, $z \wedge w = \min\{z, w\}$, $z^+ = \max\{z, 0\}$, $z^- = \max\{-z, 0\}$, and $\lfloor z \rfloor = \max\{j \in \mathbb{Z} : j \leq z\}$. Given two functions $\varphi, \check{\varphi} : \mathbb{N} \to \mathbb{R}$, we write $\check{\varphi}(n) = O(\varphi(n))$ if there exists $\kappa > 0$ and some $n_0 \in \mathbb{N}$ such that $|\check{\varphi}(n)| \leq \kappa|\varphi(n)|$ for all $n > n_0$.

CHAPTER II

BASICS OF A MULTI-SERVER QUEUE

A $G/G/n$ queue is a classic stochastic model that has been extensively studied in the literature. See, e.g., [5, 22, 30], among others. In such a system, there are n identical servers. The customer arrival process to the system is assumed to be general (the first G in the $G/G/n$ notation). Upon his arrival to the system, a customer gets into service immediately if an idle server is available. Otherwise, he waits in a buffer with infinite waiting room holding a first-in-first-out (FIFO) queue. The service times are assumed to be general (the second G), forming an arbitrary sequence of nonnegative random variables. When a server finishes serving a customer, the server takes the leading customer from the waiting buffer. When the queue is empty, the server begins to idle. A $G/G/n$ queue is also referred to as a parallel-server queue. To model customer abandonment in this model, each customer is assigned a patience time. When a customer's waiting time in the buffer exceeds his patience time, the customer abandons the system without any service. Retrial is not modeled in this thesis and we assume that a customer will not abandon the system when he is in service. If the patience times are general, the resulting model is referred to as a $G/G/n+G$ queue. If we further assume that the patience times are iid, it is referred to as a $G/G/n+GI$ queue. The interarrival times and the service times are assumed to be general, without the iid assumptions.

The mathematical definition of a $G/G/n+G$ queue is given in Section 2.1. Section 2.2 is dedicated to several preliminary results for the $G/G/n+G$ queue.

2.1 A $G/G/n+G$ queue

To define a $G/G/n+G$ queue, we are given a sequence of nonnegative primitive random variables $\{\tau_i, v_i, \gamma_i : i \in \mathbb{Z}\}$. We assume that $\tau_i(\omega) \leq \tau_{i+1}(\omega)$ for each sample path $\omega \in \Omega$ and each $i \in \mathbb{Z}$. One interprets $\tau_i(\omega)$ as the arrival time of the ith customer. We further assume that for each $\omega \in \Omega$, $\tau_1(\omega) > 0$ and $\tau_i(\omega) = 0$ for all $i \leq 0$. Thus, by time zero, all customers with indices $i \leq 0$ have arrived at the system, and $\tau_1(\omega)$ is the arrival time of the first customer after time zero. For $t \geq 0$, let

$$E(t) = \sup\{i \in \mathbb{Z}_+ : \tau_i \leq t\}. \tag{2.1}$$

Clearly, $E(t)$ is the number of customers who arrive at the system during $(0, t]$.

For each $\omega \in \Omega$ and $i \in \mathbb{Z}$, one interprets $v_i(\omega)$ as the service time of the ith customer if he has not started his service by time zero, or his remaining service time at time zero if he has started service. Let

$$N(0, \omega) = \inf\{i \geq 0 : v_j(\omega) = 0 \text{ for all } j \leq -i\}.$$

We assume that $N(0, \omega) < \infty$ on each sample path ω. The integer $N(0, \omega)$ is interpreted as the number of total customers who are in the system at time zero. Let

$$Q(0, \omega) = (N(0, \omega) - n)^+,$$

which is interpreted as the number of customers who are waiting in the buffer at time zero. Thus, customers $i = 1 - N(0, \omega), \ldots, 0$ are in the system at time zero, with customers $i = 1 - Q(0, \omega), \ldots, 0$ waiting in the buffer.

For $i \geq 1$, $\gamma_i(\omega) \geq 0$ is interpreted as the patience time of the ith customer. For customer i who is waiting in the buffer at time zero, $\gamma_i(\omega) > 0$ is interpreted as the remaining patience time of the customer. For customer i who has entered service or abandoned the system by time zero, $\gamma_i(\omega)$ can take any value. For future purposes, we set $\gamma_i(\omega) = -1$ when $i \leq -Q(0, \omega)$. To keep track of the history of the $G/G/n+G$

queue, we define a filtration $\{\mathscr{F}_i : i \in \mathbb{Z}_+\}$ by

$$\mathscr{F}_i = \sigma\{\tau_{j+1}, v_j, \gamma_j : j \le i\}. \tag{2.2}$$

Most of this thesis studies $G/G/n + GI$ queues. In such a queue, the sequence of patience times $\{\gamma_i : i \in \mathbb{N}\}$ is assumed to be iid.

2.2 Preliminary results on the $G/G/n + G$ queue

In this section, we study the $G/G/n + G$ queue where the patience times are *not* assumed to be iid. In such a queue, the interarrival times, the service times, and the patience times are three arbitrary sequences of nonnegative random variables. We first rigorously define offered waiting times and virtual waiting times. The offered waiting time of each customer is shown to be measurable in Lemma 2.1. Then, in Lemma 2.2, these times are shown to be related at the arrival time of each customer. We next define nominal service-starting times. These nominal times are shown to be ordered in the FIFO fashion in Lemma 2.3. A relationship among the offered waiting times, the patience times, and the queue length process is presented in Lemma 2.4. These lemmas will be used to prove the theorems in Chapters 3 and 4.

First, we introduce two notions: *offered waiting times* and *virtual waiting times*. See [2, 51] for discussions on them in single-server queues. In a $G/G/n + G$ queue, for each $i \in \mathbb{Z}$, we use w_i to denote the offered waiting time of the ith customer: For $i \ge 1$, w_i is the amount of time he would have to wait in the buffer until getting into service, if his patience were infinite; for $1 - Q(0) \le i \le 0$, the ith customer is waiting in the buffer at time zero and w_i is his remaining waiting time if he had infinite patience. To define an offered waiting time mathematically, it is convenient to introduce the *remaining service time process*, $r_i = \{r_i(t) : t \ge 0\}$ for $i \in \mathbb{Z}$, where $r_i(t)$ is the remaining service time of the ith customer at time t. Fix $\omega \in \Omega$. For each

$i \le -N(0,\omega)$, let $w_i(\omega) = 0$ and $r_i(t,\omega) = 0$ for all $t \ge 0$, and for $i > -N(0,\omega)$, let

$$w_i(\omega) = \inf\left\{t \ge 0 : \sum_{j \le i-1} 1_{\{r_j(\tau_i + t) > 0\}}(\omega) < n\right\} \tag{2.3}$$

and

$$r_i^{\mathrm{a}}(t,\omega) = 1_{\{t < \tau_i + \gamma_i\}}(\omega) v_i(\omega), \tag{2.4}$$

$$r_i^{\mathrm{s}}(t,\omega) = 1_{\{t < \tau_i + w_i\}}(\omega) v_i(\omega) + 1_{\{t \ge \tau_i + w_i\}}(\omega)(v_i(\omega) - t)^+, \tag{2.5}$$

$$r_i(t,\omega) = 1_{\{0 \le \gamma_i \le w_i\}}(\omega) r_i^{\mathrm{a}}(t,\omega) + (1 - 1_{\{0 \le \gamma_i \le w_i\}}(\omega)) r_i^{\mathrm{s}}(t,\omega) \tag{2.6}$$

for $t \ge 0$. Equation (2.3) says that if no arrival occurs after the $(i-1)$st customer, w_i is the amount of time beyond τ_i until one of the n servers becomes idle. Equation (2.6) says that the ith customer will abandon the queue if $0 \le \gamma_j \le w_j$, and in this case his remaining service time at time t is given by $r_i^{\mathrm{a}}(t)$; otherwise, he either has received or will receive service, and his remaining service time at time t is $r_i^{\mathrm{s}}(t)$. Clearly, recursions (2.3)–(2.6) define $w_i(\omega)$ for each $\omega \in \Omega$ and $i \in \mathbb{Z}$.

Our first lemma demonstrates the measurability of each offered waiting time.

Lemma 2.1. *For a $G/G/n{+}G$ queue, w_i is $\mathscr{F}_k$-measurable for $k \in \mathbb{Z}_+$ and $i \le k+1$, where the filtration $\{\mathscr{F}_k : k \in \mathbb{Z}\}$ is defined by (2.2).*

Proof. For each $m \in \mathbb{Z}_+$, let $v_{i,m} = 1_{\{i > -m\}} v_i$, $\gamma_{i,m} = 1_{\{i > -m\}}\gamma_i - 1_{\{i \le -m\}}$, and $\mathscr{F}_{k,m} = \sigma\{\tau_{i+1}, v_{i,m}, \gamma_{i,m} : i \le k\}$. Because $v_i = \lim_{m\to\infty} v_{i,m}$ and $\gamma_i = \lim_{m\to\infty}\gamma_{i,m}$ for all $i \in \mathbb{Z}$, we have $\mathscr{F}_k = \bigvee_{m=0}^{\infty} \mathscr{F}_{k,m}$, where $\bigvee_{m=0}^{\infty} \mathscr{F}_{k,m}$ is the smallest σ-field that contains each $\mathscr{F}_{k,m}$ for $m \in \mathbb{Z}_+$. Given $k \ge 0$ and $m \ge 0$, we define $w_{i,m}$ and $r_{i,m}(t)$ recursively via a similar procedure as in (2.3)–(2.6) for w_i and $r_i(t)$: For $i \le -m$, let $w_{i,m} = 0$ and $r_{i,m}(t) = 0$ for $t \ge 0$; for $i \ge -m+1$, let

$$w_{i,m} = \inf\left\{t \ge 0 : \sum_{j \le i-1} 1_{\{r_{j,m}(\tau_i + t) > 0\}} < n\right\} \tag{2.7}$$

and

$$r^{\mathrm{a}}_{i,m}(t) = 1_{\{t<\tau_i+\gamma_{i,m}\}} v_{i,m}, \tag{2.8}$$

$$r^{\mathrm{s}}_{i,m}(t) = 1_{\{t<\tau_i+w_{i,m}\}} v_{i,m} + 1_{\{t\geq\tau_i+w_{i,m}\}} (v_{i,m}-t)^+, \tag{2.9}$$

$$r_{i,m}(t) = 1_{\{0\leq\gamma_{i,m}\leq w_{i,m}\}} r^{\mathrm{a}}_{i,m}(t) + (1-1_{\{0\leq\gamma_{i,m}\leq w_{i,m}\}}) r^{\mathrm{s}}_{i,m}(t) \tag{2.10}$$

for $t \geq 0$. By (2.7), we get $w_{i,m} = 0$ for $i \leq -m+n$.

Fix integers $k \geq 0$ and $m \geq 0$. We would like to show that $w_{i,m}$ is $\mathscr{F}_{k,m}$-measurable for each $i \leq k+1$. Assume that there exists an integer $j \leq k$ such that $w_{i,m}$ is $\mathscr{F}_{k,m}$-measurable for all $i \leq j$. Clearly, $j = (-m+n) \wedge k$ is such a choice. To prove by induction on j that $w_{i,m}$ is $\mathscr{F}_{k,m}$-measurable, it remains to show that $w_{j+1,m}$ is also $\mathscr{F}_{k,m}$-measurable. To see this, for any $t \geq 0$, $r^{\mathrm{a}}_{i,m}(t)$, $r^{\mathrm{s}}_{i,m}(t)$, and $r_{i,m}(t)$ are $\mathscr{F}_{k,m}$-measurable. By (2.8)–(2.10), the process $r_{i,m}$ is right-continuous, and thus $r_{i,m}(\tau_i+t)$ is $\mathscr{F}_{k,m}$-measurable for $i \leq j$ and $t \geq 0$, because τ_i is $\mathscr{F}_{k,m}$-measurable. Since

$$\{w_{j+1,m} \leq t\} = \left\{ \sum_{i\leq j} 1_{\{r_{i,m}(\tau_{j+1}+t)>0\}} < n \right\},$$

we conclude that $w_{j+1,m}$ is $\mathscr{F}_{k,m}$-measurable, thus proving the measurability of $w_{i,m}$.

Given $\omega \in \Omega$, we have $v_i(\omega) = v_{i,m}(\omega)$ and $\gamma_i(\omega) = \gamma_{i,m}(\omega)$ for all $m \geq N(0,\omega)$ and $i \in \mathbb{Z}$. One can check that $w_i(\omega) = w_{i,m}(\omega)$ for $m \geq N(0,\omega)$ and thus

$$w_i = \lim_{m\to\infty} w_{i,m}.$$

Therefore, w_i is $\mathscr{F}_k$-measurable for $i \leq k+1$. □

For the $G/G/n+G$ queue, we use $W(t)$ to denote its virtual waiting time at time $t \geq 0$. One interprets $W(t)$ as the amount of time a hypothetical customer would have to wait in the buffer, had he arrived at time t with infinite patience. Given $N(0)$, the number of total customers in the system at time zero, the *virtual waiting time at time t* can be defined by

$$W(t) = \inf\left\{ s \geq 0 : \sum_{i=1-N(0)}^{E(t)} 1_{\{r_i(t+s)>0\}} < n \right\}. \tag{2.11}$$

We call $W = \{W(t) : t \geq 0\}$ the *virtual waiting time process*. The following lemma relates offered waiting times to virtual waiting times at the corresponding arrival times.

Lemma 2.2. *For a $G/G/n + G$ queue,*

$$W(\tau_i-) \leq w_i \leq W(\tau_i) \quad \textit{for } i \geq 1$$

and

$$w_i \leq W(0) \quad \textit{for } i \leq 0.$$

Proof. Let $y(t) = \inf\{s \geq 0 : \sum_{i=1-N(0)}^{E(t-)} 1_{\{r_i(s)>0\}} < n\}$. Then, for any $u \in [\tau_{E(t-)}, t)$, since $E(u) = E(t-)$, using (2.11) we have $u + W(u) = u \vee y(t)$. Thus, $W(u) = (y(t) - u)^+$ and $W(t-) = (y(t) - t)^+$. Since $E(\tau_i-) < i \leq E(\tau_i)$, it follows from (2.3) and (2.11) that $W(\tau_i-) \leq w_i \leq W(\tau_i)$; in particular, $W(\tau_i-) = w_i$ if exactly one customer arrives at time τ_i. Using $E(0) = 0$ and $\tau_i = 0$ for $i \leq 0$, $w_i \leq W(0)$ also follows from (2.3) and (2.11). □

Note that the ith customer would begin his service at time $\tau_i + w_i$ if he would not abandon the queue. We call $\tau_i + w_i$ the ith customer's *nominal service-starting time*. It follows from (2.3) that

$$\tau_i + w_i = \inf\left\{s \geq \tau_i : \sum_{j \leq i-1} 1_{\{r_j(s)>0\}} < n\right\}. \tag{2.12}$$

Similarly, we call $t + W(t)$ the nominal service-starting time for a customer arriving at time t. It can be written as

$$t + W(t) = \inf\left\{s \geq t : \sum_{i=1-N(0)}^{E(t)} 1_{\{r_i(s)>0\}} < n\right\}. \tag{2.13}$$

The lemma below states that although customer abandonment is involved, the nominal service-starting times are still ordered in the FIFO fashion as in a $G/G/n$ queue without abandonment.

Lemma 2.3. *For a $G/G/n+G$ queue,*

$$t_1 + W(t_1) \le t_2 + W(t_2) \quad \textit{for } 0 \le t_1 \le t_2,$$

and

$$\tau_i + w_i \le \tau_j + w_j \quad \textit{for any } i, j \in \mathbb{Z} \textit{ with } i \le j.$$

Proof. By (2.4)–(2.6), the process $\{1_{\{r_i(t)>0\}} : t \ge 0\}$ is right-continuous for all $i \in \mathbb{Z}$. Hence,

$$\sum_{i=1-N(0)}^{E(t)} 1_{\{r_i(t+W(t))>0\}} < n.$$

If there exist $0 \le t_1 \le t_2$ such that $t_1 + W(t_1) > t_2 + W(t_2)$,

$$n \le \sum_{i=1-N(0)}^{E(t_1)} 1_{\{r_i(t_2+W(t_2))>0\}} \le \sum_{i=1-N(0)}^{E(t_2)} 1_{\{r_i(t_2+W(t_2))>0\}} < n$$

by (2.13), which yields a contradiction. Thus, $t_1 + W(t_1) \le t_2 + W(t_2)$. Using (2.12) we can prove $\tau_i + w_i \le \tau_j + w_j$ for $i \le j$ by a similar argument. □

Before getting into service or abandoning the queue, the ith customer waits $\gamma_i \wedge w_i$ units of time in the buffer. If the ith customer arrived at the system before time t and $t < \tau_i + (\gamma_i \wedge w_i)$, he must be waiting in the buffer at time t. Hence, the number of waiting customers can be counted by

$$Q(t) = \sum_{i=1-N(0)}^{E(t)} 1_{\{t<\tau_i+(\gamma_i \wedge w_i)\}}.$$

The process $Q = \{Q(t) : t \ge 0\}$ is called the *queue length process*.

The last lemma of this section establishes a pair of inequalities. The inequalities will later allow us to convert a summation of offered waiting times into an integral of the queue length process.

Lemma 2.4. *For a $G/G/n+G$ queue,*

$$\int_0^t Q(s)\,\mathrm{d}s \le \sum_{i=1-Q(0)}^{E(t)} (\gamma_i \wedge w_i) \le \int_0^{t+W(t)} Q(s)\,\mathrm{d}s \quad \textit{for all } t \ge 0.$$

Proof. For $i \geq 1 - Q(0)$, the ith customer spends $\gamma_i \wedge w_i$ units of time waiting in the buffer. For $t \geq 0$, let

$$b_i(t) = 1_{\{1-N(0) \leq i \leq E(t),\, t < \tau_i + (\gamma_i \wedge w_i)\}}.$$

Then, $b_i(t) = 1$ if the ith customer is waiting in the buffer at time t and $b_i(t) = 0$ otherwise. As a result,

$$q_i(t) = \int_0^t b_i(s)\,\mathrm{d}s$$

is the ith customer's cumulative waiting time by t. Note that $q_i(t) \leq \gamma_i \wedge w_i$, and $q_i(t) = \gamma_i \wedge w_i$ holds if and only if the ith customer has got service or abandoned the queue by t. For any $0 \leq s \leq t$, the queue length at time s can be counted by $Q(s) = \sum_{i=1-Q(0)}^{E(t)} b_i(s)$. Then,

$$\int_0^t Q(s)\,\mathrm{d}s = \sum_{i=1-Q(0)}^{E(t)} \int_0^t b_i(s)\,\mathrm{d}s = \sum_{i=1-Q(0)}^{E(t)} q_i(t) \leq \sum_{i=1-Q(0)}^{E(t)} (\gamma_i \wedge w_i).$$

For $1 - Q(0) \leq i \leq E(t)$, the ith customer should have got into service or abandoned the system by time $t + W(t)$, because $\tau_i + w_i \leq t + W(t)$ (see Lemmas 2.2 and 2.3). Then $q_i(t + W(t)) = \gamma_i \wedge w_i$. It follows that

$$\int_0^{t+W(t)} Q(s)\,\mathrm{d}s = \sum_{i=1-Q(0)}^{E(t+W(t))} q_i(t + W(t)) \geq \sum_{i=1-Q(0)}^{E(t)} q_i(t + W(t)) = \sum_{i=1-Q(0)}^{E(t)} (\gamma_i \wedge w_i).$$

□

CHAPTER III

AN ASYMPTOTIC RELATIONSHIP

The focus of this chapter is the $G/G/n+GI$ model whose patience times are iid following a general distribution. Let $A(t)$ be the cumulative number of customers who have abandoned the system by time t. The purpose of this chapter is to establish an asymptotic relationship between the queue length process $Q = \{Q(t) : t \geq 0\}$ and the *abandonment process* $A = \{A(t) : t \geq 0\}$ in a $G/G/n+GI$ queue when the number of servers n is large. This relationship plays a crucial role in proving the limit theorems in Chapter 4.

To motivate this relationship, consider an $M/M/n+M$ queue in which the sequences of interarrival times, service times and patience times are all iid and each sequence follows an exponential distribution. Each customer waiting in the buffer abandons the system at rate $\alpha \geq 0$. Because of the memoryless property of an exponential distribution, one can argue that with probability one,

$$A(t) = N_{\mathrm{P}}\left(\alpha \int_0^t Q(s)\,\mathrm{d}s\right) \quad \text{for all } t \geq 0, \tag{3.1}$$

where $N_{\mathrm{P}} = \{N_{\mathrm{P}}(t) : t \geq 0\}$ is a Poisson process with unity rate.

To further simplify relationship (3.1), let us consider a sequence of $M/M/n+M$ queues indexed by the number of servers n. For the nth system, let $A^n(t)$ be the number of abandonments by time t and $Q^n(t)$ be the queue length at time t. Suppose that these queues are operated in the QED regime, i.e., condition (1.1) holds. One can prove from (3.1) that for each $T > 0$,

$$\frac{1}{\sqrt{n}} \sup_{0 \leq t \leq T} \left| A^n(t) - \alpha \int_0^t Q^n(s)\,\mathrm{d}s \right| \to 0 \quad \text{in probability as } n \to \infty,$$

i.e., the asymptotic relationship (1.2) holds. The main result of this chapter is to prove

that (1.2) holds for a sequence of $G/G/n+GI$ queues with general arrival processes that can be time-nonhomogeneous, assuming that the sequence of diffusion-scaled queue length processes is stochastically bounded.

The heavy-traffic condition (1.1) implies that the sequence of queues is critically loaded in the limit. It is often used to prove stochastic boundedness for the diffusion-scaled queue length processes. However, condition (1.1) is not necessary for the stochastic boundedness result. For example, when the sequence of $M/M/n+M$ systems is underloaded, namely, $\lim_{n\to\infty}\rho^n < 1$, the stochastic boundedness still holds. Theorem 3.1 assumes the stochastic boundedness for the sequence of diffusion-scaled queue length processes. The heavy-traffic condition (1.1) is *not* used in the rest of this chapter. For a particular sequence of $G/G/n+G$ systems, proving the stochastic boundedness result is by no means easy. The second theorem of this chapter is a comparison result showing that the queue length at any time in a $G/G/n+G$ queue is dominated by the queue length in the corresponding $G/G/n$ queue with longer service times and no customer abandonment. The comparison result implies that it is sufficient to prove the stochastic boundedness for the diffusion-scaled queue length processes in a sequence of $G/G/n$ queues without customer abandonment.

In Theorem 3.1, α in (1.2) the density (as the right derivative) at zero of the patience time distribution. Under the stochastic boundedness assumption on the diffusion-scaled queue length processes, the customer waiting times will be proved to converge to zero as $n\to\infty$. Thus, customer abandonment rarely happens when n is large. Only those customers who have extremely small patience times can possibly abandon the system. Therefore, the patience time distribution, outside a small neighborhood of zero, barely has any influence on the system dynamics.

In Section 3.1, we introduce a sequence of $G/G/n+GI$ queues and the stochastic boundedness assumption on the diffusion-scaled queue length processes. The main results of this chapter, including the asymptotic relationship and the comparison

result, are presented in Section 3.2. The detailed proof of the asymptotic relationship is given in Section 3.3. The proof of Theorem 3.2 can be found in Section 3.4. We discuss an initial condition of Theorem 3.1 in Section 3.5.

3.1 Asymptotic framework on $G/G/n+GI$ queues

Consider a sequence of $G/G/n+GI$ queues indexed by the number of servers n. We add a superscript n to the primitive random variables of the nth system and use $\{\mathscr{F}_i^n : i \in \mathbb{Z}_+\}$ to denote the associated filtration, given by

$$\mathscr{F}_i^n = \sigma\{\tau_{j+1}^n, v_j^n, \gamma_j^n : j \leq i\}. \tag{3.2}$$

We assume that

$$\gamma_{i+1}^n \text{ is independent of } \mathscr{F}_i^n \text{ for each } i \in \mathbb{Z}_+ \tag{3.3}$$

and that $\{\gamma_i^n : i \in \mathbb{N}\}$ is a sequence of iid random variables with distribution function F that does not change with n. Recall that for $i \geq 1$, γ_i^n is the patience time of the ith customer who arrives after time zero at the nth system. The preceding assumption states that the distribution of these patience times does not depend on the number of servers, which is reasonable in many cases. For $i \leq 0$, γ_i^n is the remaining patience time of a customer who is waiting in the buffer of the nth system at time zero. This remaining patience time may depend on how long the customer has been waiting by time zero, and this waiting time may in turn depend on the number of servers n. We further assume that the distribution F satisfies

$$F(0) = 0 \tag{3.4}$$

and is right-differentiable at zero with right derivative

$$\alpha = \lim_{t \downarrow 0} t^{-1} F(t) < \infty. \tag{3.5}$$

The customer arrival process of the nth system is $E^n = \{E^n(t) : t \geq 0\}$, where $E^n(t)$, defined in (2.1), denotes the number of customer arrivals in $(0, t]$. The

fluid-scaled arrival process $\bar{E}^n$ is defined by

$$\bar{E}^n(t) = \frac{1}{n} E^n(t).$$

The following two assumptions are made upon the arrival processes. First, given an arbitrary $T > 0$, there exists a constant $c_T > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}\Big[\inf_{0\le t\le T} \{\bar{E}^n(t+\delta) - \bar{E}^n(t)\} < \delta c_T \Big] = 0 \quad \text{for all } \delta > 0. \tag{3.6}$$

Moreover, the sequence of fluid-scaled arrival processes is stochastically bounded, i.e., for each $T > 0$,

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}[\bar{E}^n(T) > a] = 0. \tag{3.7}$$

Roughly speaking, condition (3.6) states that when n is large, the number of customer arrivals should be at least $n\delta c_T$ during $(t, t+\delta]$ for any $t \in [0, T]$; in other words, the arrival rate of the nth system is in the order of $O(n)$. Assumptions (3.6) and (3.7) impose very mild constraints on the arrival processes. Clearly, they allow each arrival process E^n to have a time-nonhomogeneous arrival rate.

Recall the queue length process Q^n and the abandonment process A^n of the nth system. We define their respective diffusion-scaled versions $\tilde{Q}^n$ and $\tilde{A}^n$ via

$$\tilde{Q}^n(t) = \frac{1}{\sqrt{n}} Q^n(t) \quad \text{and} \quad \tilde{A}^n(t) = \frac{1}{\sqrt{n}} A^n(t).$$

A key assumption of Theorem 3.1 is that the sequence of diffusion-scaled queue length processes is stochastically bounded, namely, for each $T > 0$,

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}\Big[\sup_{0\le t\le T} \tilde{Q}^n(t) > a \Big] = 0. \tag{3.8}$$

Roughly speaking, it requires that Q^n be in the order of $O(n^{1/2})$.

For Theorem 3.1, we also need to make an assumption on the initial condition. Let $G^n(0)$ be the number of customers who are waiting in the buffer at time zero but will eventually abandon the system. Let

$$\tilde{G}^n(0) = \frac{1}{\sqrt{n}} G^n(0).$$

We assume that

$$\tilde{G}^n(0) \Rightarrow 0 \quad \text{as } n \to \infty. \tag{3.9}$$

Clearly, condition (3.9) is satisfied if no customers are waiting in the buffer at time zero. The validity of this initial condition will be further discussed in Section 3.5.

3.2 The asymptotic relationship

We state two theorems in this section. The first theorem is the main result of this chapter. It says that the relationship (1.2) holds for a sequence of $G/G/n+GI$ queues under certain conditions.

Theorem 3.1. *Consider a sequence of $G/G/n+GI$ queues that satisfies (3.3)–(3.7). Assume that the sequence of diffusion-scaled queue length processes is stochastically bounded and the sequence of queues satisfies the initial condition (3.9). Then, the asymptotic relationship (1.2) holds for each $T > 0$.*

The proof of Theorem 3.1 is presented in Section 3.3. All assumptions in Theorem 3.1 are standard, except for the stochastic boundedness assumption (3.8). Verifying this assumption can be a significant task.

We now present the second theorem. This theorem, referred to as the comparison result, states that the queue length at any time in a $G/G/n+G$ queue is dominated by the queue length in the corresponding $G/G/n$ queue with longer service times and no customer abandonment. This comparison result implies that, to verify the stochastic boundedness assumption (3.8) for a sequence of $G/G/n+GI$ queues, it is sufficient to prove stochastic boundedness for the queue length processes in the corresponding $G/G/n$ queues.

To state Theorem 3.2, we consider two FIFO queues: a $G/G/n+G$ queue denoted by $\Sigma^{(1)}$ and a $G/G/n$ queue denoted by $\Sigma^{(2)}$. For $\ell = 1, 2$, we add a superscript (ℓ) to the primitive random variables and performance processes of $\Sigma^{(\ell)}$. We assume that all servers in both systems are identical, the arrival processes to both queues

are identical, and at time zero, there are $N(0)$ customers in each system, indexed by $i = 1 - N(0), \ldots, 0$. Recall that $v_i^{(\ell)}$ is the service time of the ith customer if he has not started service by time zero, or his remaining service time at time zero if he has started service. We further assume that

$$v_i^{(1)} \leq v_i^{(2)} \quad \text{for all } i \geq 1 - N(0). \tag{3.10}$$

In short, there are two differences between the queues. First, each customer in $\Sigma^{(1)}$ have an equal or shorter service time than the corresponding customer in $\Sigma^{(2)}$. Second, customers in $\Sigma^{(1)}$ can possibly abandon the system whereas those in $\Sigma^{(2)}$ cannot.

Theorem 3.2. *Let $Q^{(1)}(t)$ and $Q^{(2)}(t)$ be the respective numbers of customers waiting in the buffers of $\Sigma^{(1)}$ and $\Sigma^{(2)}$ at time $t \geq 0$. Then, on each sample path,*

$$Q^{(1)}(t) \leq Q^{(2)}(t) \quad \textit{for all } t \geq 0.$$

The proof of Theorem 3.2 is given in Section 3.4

3.3 Proof of Theorem 3.1

We present the proof of Theorem 3.1 in this section. The proof is decomposed into three propositions, which are proved in Sections 3.3.2–3.3.4.

Our attention is now focused on a sequence of $G/G/n + GI$ queues that satisfies conditions (3.3)–(3.9). Let $G^n(t)$ denote, among all customers who have arrived at the nth system by time $t \geq 0$, the number of those who will eventually abandon the queue. The process $G^n = \{G^n(t) : t \geq 0\}$ has a diffusion-scaled version

$$\tilde{G}^n(t) = \frac{1}{\sqrt{n}} G^n(t).$$

Our first result is the following proposition showing that A^n and G^n are asymptotically close in diffusion scaling.

Proposition 3.3. *Under the conditions of Theorem 3.1,*

$$\tilde{A}^n - \tilde{G}^n \Rightarrow 0 \quad \textit{as } n \to \infty.$$

The proof of the proposition is presented in Section 3.3.4. To prove Theorem 3.1 given Proposition 3.3, it suffices to show that for each $T > 0$,

$$\sup_{0\le t\le T}\left|\tilde{G}^n(t) - \alpha\int_0^t \tilde{Q}^n(s)\,\mathrm{d}s\right| \to 0 \quad \text{in probability as } n\to\infty. \tag{3.11}$$

To prove (3.11), one needs to further analyze the process G^n. For the nth system, we use $W^n(t)$ and w_i^n to denote the corresponding virtual and offered waiting times. For each customer $i \ge 1 - Q^n(0)$, given his patience time γ_i^n and offered waiting time w_i^n, one can determine whether the customer will eventually abandon the queue. He will wait γ_i^n units of time and leave the system with no service when $\gamma_i^n \le w_i^n$, or will wait w_i^n units of time and get into a server otherwise. This implies the following expression

$$G^n(t) = \sum_{i=1-Q^n(0)}^{E^n(t)} 1_{\{\gamma_i^n \le w_i^n\}}.$$

Clearly, the process G^n can be decomposed into

$$G^n(t) = G^n(0) + G_1^n(t) + G_2^n(t), \tag{3.12}$$

where

$$G^n(0) = \sum_{i=1-Q^n(0)}^{0} 1_{\{\gamma_i^n \le w_i^n\}}$$

is the number of customers who are initially waiting in the buffer but will eventually abandon the queue,

$$G_1^n(t) = \sum_{i=1}^{E^n(t)} \left(1_{\{\gamma_i^n \le w_i^n\}} - F\left(w_i^n\right)\right) \quad \text{and} \quad G_2^n(t) = \sum_{i=1}^{E^n(t)} F(w_i^n).$$

Defining the diffusion-scaled versions

$$\tilde{G}_1^n(t) = \frac{1}{\sqrt{n}} G_1^n(t) \quad \text{and} \quad \tilde{G}_2^n(t) = \frac{1}{\sqrt{n}} G_2^n(t),$$

we have the following two propositions.

Proposition 3.4. *Under the conditions of Theorem 3.1,*

$$\tilde{G}_1^n \Rightarrow 0 \quad \textit{as } n\to\infty.$$

Proposition 3.5. *Under the conditions of Theorem 3.1, for any $T > 0$,*

$$\sup_{0\le t\le T}\left|\tilde{G}_2^n(t) - \alpha\int_0^t \tilde{Q}^n(s)\,\mathrm{d}s\right| \Rightarrow 0 \quad \textit{as } n\to\infty.$$

The proofs of Propositions 3.4 and 3.5 are presented in Sections 3.3.2 and 3.3.3, respectively. Clearly, the proof of Theorem 3.1 follows from (3.9), (3.12) and Propositions 3.3–3.5.

3.3.1 Virtual waiting time process

This section is a preparation for proving Propositions 3.3–3.5. The main result here is Proposition 3.6, which says that for the sequence of $G/G/n+GI$ queues, the virtual waiting time processes converges to zero in probability.

Proposition 3.6. *Assume that (3.4) and (3.6)–(3.8) hold. Then,*

$$W^n \Rightarrow 0 \quad \textit{as } n\to\infty.$$

Intuitively, this convergence follows from the following observation. Fix $t > 0$. By assumption (3.6), for any $\delta > 0$ small enough, there exists a constant $\kappa > 0$ such that when n is large, there are at least $n\kappa\delta$ customer arrivals during $(t, t+\delta]$. When n is large, all customers who arrived before time t must have entered service or abandoned the system by time $t+\delta$. Otherwise, except for a small abandoning portion, those who arrived during $(t, t+\delta]$ must reside in the buffer at time $t+\delta$ because of the FIFO discipline, contradicting the stochastic boundedness assumption on the diffusion-scaled queue length processes. Therefore, the virtual waiting time $W^n(t)$ should be no more than δ. Since $\delta > 0$ is arbitrary, this implies that $W^n(t)$ goes to zero as n goes to infinity.

Before presenting the proof of Proposition 3.6, we give a few corollaries that will be used in later proofs. Define $g_F : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$g_F(t) = \begin{cases} \alpha & \text{for } t = 0, \\ t^{-1}F(t) & \text{for } t > 0. \end{cases}$$

Under assumptions (3.4) and (3.5), g_F is right-continuous at zero and

$$F(t) = t g_F(t) \quad \text{for all } t \geq 0.$$

Corollary 3.7. *Assume that (3.4)–(3.8) hold. Then, for each $T > 0$,*

$$\sup_{1 \leq i \leq E^n(T)} w_i^n \Rightarrow 0, \tag{3.13}$$

$$\sup_{1 \leq i \leq E^n(T)} F(w_i^n) \Rightarrow 0, \tag{3.14}$$

$$\sup_{1 \leq i \leq E^n(T)} |g_F(w_i^n) - \alpha| \Rightarrow 0, \tag{3.15}$$

$$\sup_{1 \leq i \leq nT} F(w_i^n) \Rightarrow 0, \tag{3.16}$$

as $n \to \infty$.

Proof. First, the convergence (3.13) follows from Lemma 2.2 and Proposition 3.6. Since F is nondecreasing and right-continuous at zero, by the continuous-mapping theorem,

$$\sup_{1 \leq i \leq E^n(T)} F(w_i^n) \leq F\left(\sup_{1 \leq i \leq E^n(T)} w_i^n \right) \Rightarrow F(0) = 0,$$

which proves (3.14). For any $\varepsilon > 0$, since g_F is right-continuous at zero, there exists $\delta > 0$ such that $|g_F(t) - \alpha| \leq \varepsilon$ for all $0 \leq t \leq \delta$, and thus (3.15) follows from

$$\limsup_{n \to \infty} \mathbb{P}\Big[\sup_{1 \leq i \leq E^n(T)} |g_F(w_i^n) - \alpha| > \varepsilon \Big] \leq \limsup_{n \to \infty} \mathbb{P}\Big[\sup_{1 \leq i \leq E^n(T)} w_i^n > \delta \Big] = 0.$$

Also for any $\varepsilon > 0$,

$$\mathbb{P}\Big[\sup_{1 \leq i \leq nT} F(w_i^n) > \varepsilon \Big] \leq \mathbb{P}\Big[\sup_{1 \leq i \leq E^n(c_T^{-1} T)} F(w_i^n) > \varepsilon \Big] + \mathbb{P}[\bar{E}^n(c_T^{-1} T) < T],$$

where $c_T > 0$ is the constant given in (3.6). Then, the convergence (3.16) follows from (3.6) and (3.14). □

To prove Proposition 3.6, we introduce the following processes. For $\delta > 0$, let

$$L_\delta^n(t) = \sum_{i=1}^{\lfloor nt \rfloor} (1_{\{\gamma_i^n \leq \delta\}} - F(\delta)) \quad \text{and} \quad \bar{L}_\delta^n(t) = \frac{1}{n} L_\delta^n(t). \tag{3.17}$$

Since the sequence of patience times $\{\gamma_i^n : i \in \mathbb{N}\}$ are iid, the functional law of large numbers (FLLN) suggests

$$\bar{L}_\delta^n \Rightarrow 0 \quad \text{as } n \to \infty. \tag{3.18}$$

For each $\delta > 0$, we also define

$$G_\delta^n(t) = \sum_{i=E^n(t)+1}^{E^n(t+\delta)} 1_{\{\gamma_i^n \le \delta\}} \tag{3.19}$$

that counts the number of customers who arrive at the nth system during $(t, t+\delta]$ but whose patience times are no more than δ. It has a fluid-scaled version given by

$$\bar{G}_\delta^n(t) = \frac{1}{n} G_\delta^n(t).$$

We further introduce the fluid-scaled queue length process $\bar{Q}^n$, given by

$$\bar{Q}^n(t) = \frac{1}{n} Q^n(t),$$

which, by the stochastic boundedness assumption (3.8), satisfies

$$\bar{Q}^n \Rightarrow 0 \quad \text{as } n \to \infty. \tag{3.20}$$

Proof of Proposition 3.6. We first claim that when $0 < \delta < W^n(t)$,

$$E^n(t+\delta) - E^n(t) \le Q^n(t+\delta) + G_\delta^n(t). \tag{3.21}$$

To see (3.21), fix $t \ge 0$. For $\delta \in (0, W^n(t))$ and $\tau_i^n \in (t, t+\delta]$, since

$$t + \delta < t + W^n(t) \le \tau_i^n + w_i^n$$

(see Lemmas 2.2 and 2.3), the ith customer will not get into service by time $t+\delta$, so he will either be waiting in queue or have abandoned the system by then. The latter case implies $\gamma_i^n < \delta$. This proves (3.21). Assume that $\delta > 0$ is small enough so that $F(\delta) < 1/2$. For each $T > 0$, inequality (3.21) implies that

$$\mathbb{P}\Big[\sup_{0 \le t \le T} W^n(t) > \delta\Big] \le \mathbb{P}\Big[\inf_{0 \le t \le T} \{E^n(t+\delta) - E^n(t) - G_\delta^n(t) - Q^n(t+\delta)\} \le 0\Big].$$

By (3.17) and (3.19),

$$G_\delta^n(t) = F(\delta)(E^n(t+\delta) - E^n(t)) + L_\delta^n(\bar{E}^n(t+\delta)) - L_\delta^n(\bar{E}^n(t)),$$

and thus

$$E^n(t+\delta) - E^n(t) - G_\delta^n(t) \geq \frac{1}{2}(E^n(t+\delta) - E^n(t)) - L_\delta^n(\bar{E}^n(t+\delta)) + L_\delta^n(\bar{E}^n(t)).$$

Then, we have

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{0\leq t\leq T} W^n(t) > \delta\Big] \\
&\leq \mathbb{P}\Big[\inf_{0\leq t\leq T}\Big\{\frac{1}{2}(\bar{E}^n(t+\delta) - \bar{E}^n(t)) - \bar{L}_\delta^n(\bar{E}^n(t+\delta)) + \bar{L}_\delta^n(\bar{E}^n(t)) - \bar{Q}^n(t+\delta)\Big\} \leq 0\Big] \\
&\quad \leq \mathbb{P}\Big[\inf_{0\leq t\leq T}\{\bar{E}^n(t+\delta) - \bar{E}^n(t)\} \leq \frac{\delta c_T}{2}\Big] + \mathbb{P}\Big[\sup_{0\leq t\leq T}|\bar{L}_\delta^n(\bar{E}^n(t+\delta))| \geq \frac{\delta c_T}{12}\Big] \\
&\qquad + \mathbb{P}\Big[\sup_{0\leq t\leq T}|\bar{L}_\delta^n(\bar{E}^n(t))| \geq \frac{\delta c_T}{12}\Big] + \mathbb{P}\Big[\sup_{0\leq t\leq T}\bar{Q}^n(t+\delta) \geq \frac{\delta c_T}{12}\Big].
\end{aligned}$$

By (3.7) and (3.18), we see $\bar{L}_\delta^n \circ \bar{E}^n \Rightarrow 0$ as $n \to \infty$. This, together with (3.6) and (3.20), yields $W^n \Rightarrow 0$. □

3.3.2 Proof of Proposition 3.4

This section is dedicated to the proof of Proposition 3.4. First, we define a continuous-time filtration $\{\mathscr{F}^n(t) : t \geq 0\}$ by

$$\mathscr{F}^n(t) = \mathscr{F}^n_{\lfloor nt \rfloor},$$

where the filtration $\{\mathscr{F}_i^n : i \in \mathbb{Z}_+\}$ is defined by (3.2). Next, let

$$H_i^n = \sum_{j=1}^{i}(1_{\{\gamma_j^n \leq w_j^n\}} - F(w_j^n))\phi(w_j^n) \quad \text{for each } i \in \mathbb{Z}_+,$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a Borel measurable function such that $0 \leq \phi(t) \leq 1$ for all $t \geq 0$. We further let

$$H^n(t) = H^n_{\lfloor nt \rfloor} \quad \text{and} \quad \tilde{H}^n(t) = \frac{1}{\sqrt{n}}H^n(t).$$

Now we establish a series of results on the process H^n.

Lemma 3.1. *Assume that (3.3) holds. Then,* $\{(H_i^n, \mathscr{F}_i^n) : i \in \mathbb{Z}_+\}$ *is a martingale.*

Proof. Lemma 2.1 ensures that w_j^n is $\mathscr{F}_i^n$-measurable for $1 \le j \le i+1$. Then, H_i^n is $\mathscr{F}_i^n$-measurable. Since w_i^n is $\mathscr{F}_{i-1}^n$-measurable whereas γ_i^n is independent of $\mathscr{F}_{i-1}^n$,

$$\mathbb{E}[H_i^n - H_{i-1}^n | \mathscr{F}_{i-1}^n] = \mathbb{E}[1_{\{\gamma_i^n \le w_i^n\}} | \mathscr{F}_{i-1}^n]\phi(w_i^n) - F(w_i^n)\phi(w_i^n) = 0.$$

Also, we have $\mathbb{E}[|H_i^n|] \le i$. So that $\{(H_i^n, \mathscr{F}_i^n) : i \in \mathbb{Z}_+\}$ is a martingale. □

Lemma 3.2. *Assume that (3.3) holds. Then,* $\{(H^n(t), \mathscr{F}^n(t)) : t \ge 0\}$ *is a martingale with quadratic variation*

$$[H^n](t) = \sum_{i=1}^{\lfloor nt \rfloor} (1_{\{\gamma_i^n \le w_i^n\}} - F(w_i^n))^2 \phi(w_i^n)^2. \tag{3.22}$$

Proof. By Lemma 3.1, H^n is adapted to $\{\mathscr{F}^n(t) : t \ge 0\}$. It is a martingale because for $0 \le s \le t$, $\mathbb{E}[|H^n(t)|] = \mathbb{E}[|H^n_{\lfloor nt \rfloor}|] < \infty$ and

$$\mathbb{E}[H^n(t) | \mathscr{F}^n(s)] = \mathbb{E}[H^n_{\lfloor nt \rfloor} | \mathscr{F}^n_{\lfloor ns \rfloor}] = H^n_{\lfloor ns \rfloor} = H^n(s).$$

Since H^n is piecewise constant and

$$\sum_{i=1}^{\lfloor nt \rfloor} |1_{\{\gamma_i^n \le w_i^n\}} - F(w_i^n)|\phi(w_i^n) \le nt,$$

the sample paths of H^n are of finite variation, from which (3.22) follows (see, e.g., Theorem 2.26 of [42]). □

Lemma 3.3. *Assume that (3.3), (3.4), and (3.6)–(3.8) hold. Then,*

$$\tilde{H}^n \Rightarrow 0 \quad \text{as } n \to \infty.$$

Before proving Lemma 3.3, we introduce a martingale convergence lemma, which is a degenerate case of the martingale FCLT. Its proof can be found in [59].

Lemma 3.4. *Let* $\{(M^n(t), \mathscr{G}^n(t)) : t \ge 0\}$ *be a local martingale with* $M^n(0) = 0$ *for each* $n \in \mathbb{N}$. *Assume that for any* $T > 0$,

$$\mathbb{E}\Big[\sup_{0<t\le T} |M^n(t) - M^n(t-)|\Big] \to 0 \quad \text{and} \quad [M^n](T) \Rightarrow 0 \quad \text{as } n \to \infty.$$

Then, $M^n \Rightarrow 0$ *as* $n \to \infty$.

Proof of Lemma 3.3. Using Lemma 3.2, $\{(\tilde{H}^n(t), \mathscr{F}^n(t)) : t \geq 0\}$ is a martingale with quadratic variation

$$[\tilde{H}^n](t) = \frac{1}{n}\sum_{i=1}^{\lfloor nt \rfloor}(1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n))^2\phi(w_i^n)^2.$$

Fix $T > 0$. By (3.16) and the fact that $F(w_i^n) \leq 1$, we obtain

$$\lim_{n\to\infty} \mathbb{E}\Big[\sup_{1\leq i\leq nT} F(w_i^n)\Big] = 0.$$

Since γ_i^n is independent of $\mathscr{F}_{i-1}^n$ but w_i^n is $\mathscr{F}_{i-1}^n$-measurable (see Lemma 2.1),

$$\mathbb{E}[(1_{\{\gamma_i^n \leq w_i^n\}} - F(w_i^n))^2\phi(w_i^n)^2|\mathscr{F}_{i-1}^n] = (1 - F(w_i^n))F(w_i^n)\phi(w_i^n)^2 \leq F(w_i^n).$$

Then,

$$\mathbb{E}[[\tilde{H}^n](T)] \leq \frac{1}{n}\sum_{i=1}^{\lfloor nT \rfloor}\mathbb{E}[F(w_i^n)] \leq T\mathbb{E}\Big[\sup_{1\leq i\leq nT} F(w_i^n)\Big].$$

It follows that $\mathbb{E}[[\tilde{H}^n](T)] \to 0$ and hence $[\tilde{H}^n](T) \Rightarrow 0$ as $n \to \infty$. Since

$$\sup_{0<t\leq T} |\tilde{H}^n(t) - \tilde{H}^n(t-)| \leq n^{-1/2},$$

we have

$$\lim_{n\to\infty} \mathbb{E}\Big[\sup_{0<t\leq T} |\tilde{H}^n(t) - \tilde{H}^n(t-)|\Big] = 0.$$

It follows from Lemma 3.4 that $\tilde{H}^n \Rightarrow 0$ as $n \to \infty$. □

For the proof of Proposition 3.4, let

$$H_1^n(t) = \sum_{j=1}^{\lfloor nt \rfloor}(1_{\{\gamma_j^n \leq w_j^n\}} - F(w_j^n)) \quad \text{and} \quad \tilde{H}_1^n(t) = \frac{1}{\sqrt{n}}H_1^n(t).$$

Proof of Proposition 3.4. Note that $H_1^n(t) = H^n(t)$ if $\phi = 1$. Lemma 3.3 implies that $\tilde{H}_1^n \Rightarrow 0$ as $n \to \infty$. Since $\tilde{G}_1^n(t) = \tilde{H}_1^n(\bar{E}^n(t))$, it follows from (3.7) that $\tilde{G}_1^n \Rightarrow 0$ as $n \to \infty$. □

3.3.3 Proof of Proposition 3.5

We present the proof of Proposition 3.5 in this section. The crucial step is using Lemma 2.4 to establish the following lemma that converts a summation of offered waiting times to an integral of the queue length process.

Lemma 3.5. *Assume that (3.3)–(3.8) hold. Then, for each $T > 0$,*

$$\sup_{0\le t\le T}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{E^n(t)} w_i^n - \int_0^t \tilde{Q}^n(s)\,\mathrm{d}s\right| \Rightarrow 0 \quad \text{as } n\to\infty.$$

Assuming Lemma 3.5, we now provide the proof of Proposition 3.5.

Proof of Proposition 3.5. We decompose $\tilde{G}_2^n(t)$ into

$$\tilde{G}_2^n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{E^n(t)} F(w_i^n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{E^n(t)} w_i^n g_F(w_i^n) = \tilde{G}_{21}^n(t) + \tilde{G}_{22}^n(t) + \alpha\int_0^t \tilde{Q}^n(s)\,\mathrm{d}s,$$

where

$$\tilde{G}_{21}^n(t) = \frac{\alpha}{\sqrt{n}}\sum_{i=1}^{E^n(t)} w_i^n - \alpha\int_0^t \tilde{Q}^n(s)\,\mathrm{d}s \quad \text{and} \quad \tilde{G}_{22}(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{E^n(t)} (g_F(w_i^n)-\alpha)w_i^n.$$

Lemma 3.5 implies that $\tilde{G}_{21}^n \Rightarrow 0$. Also by (3.8) and Lemma 3.5,

$$\lim_{a\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{E^n(T)} w_i^n > a\Big] = 0.$$

Then, it follows from (3.15) that

$$\sup_{0\le t\le T}|\tilde{G}_{22}(t)| \le \frac{1}{\sqrt{n}}\sum_{i=1}^{E^n(T)} w_i^n \sup_{1\le i\le E^n(T)}|g_F(w_i^n)-\alpha| \Rightarrow 0,$$

which concludes the proof. □

It remains to prove Lemma 3.5. Let

$$H_F^n(t) = \sum_{i=1}^{\lfloor nt\rfloor}(1_{\{\gamma_i^n\le w_i^n\}} - F(w_i^n))F(w_i^n) \quad \text{and} \quad \tilde{H}_F^n(t) = \frac{1}{\sqrt{n}}H_F^n(t).$$

Since $H_F^n(t) = H^n(t)$ if $\phi = F$, Lemma 3.3 implies that

$$\tilde{H}_F^n \Rightarrow 0 \quad \text{as } n\to\infty. \tag{3.23}$$

In the next lemma, we demonstrate the stochastic boundedness of the process $\tilde{G}_2$.

Lemma 3.6. *Assume that (3.3)–(3.8) hold. Then, for any $T > 0$*

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}[\tilde{G}_2^n(T) > a] = 0. \tag{3.24}$$

Proof. Fix $T > 0$. For $t \geq 0$, we decompose $\tilde{G}_2^n(t)$ into

$$\tilde{G}_2^n(t) = \tilde{G}_{23}^n(t) + \tilde{G}_{24}^n(t),$$

where

$$\tilde{G}_{23}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (1 - F(w_i^n))F(w_i^n) \quad \text{and} \quad \tilde{G}_{24}^n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} F(w_i^n)^2.$$

Then, the stochastic boundedness result (3.24) holds if we have

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}[\tilde{G}_{23}^n(T) > a] = 0, \tag{3.25}$$

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}[\tilde{G}_{24}^n(T) > a] = 0. \tag{3.26}$$

Using Lemma 2.4, we get

$$\sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n \leq \sum_{i=1-Q^n(0)}^{E^n(t)} (\gamma_i^n \wedge w_i^n) \leq \int_0^{t+W^n(t)} Q^n(s)\,\mathrm{d}s.$$

It follows from (3.8) and Proposition 3.6 that

$$\lim_{a\to\infty} \limsup_{n\to\infty} \mathbb{P}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n > a\right] = 0. \tag{3.27}$$

Note that

$$\tilde{G}_{23}^n(t) - \frac{\alpha}{\sqrt{n}} \sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n = \tilde{H}_F^n(\bar{E}^n(t)) + \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n (g_F(w_i^n) - \alpha). \tag{3.28}$$

The first term on the right side of (3.28) satisfies $\tilde{H}_F^n \circ \bar{E}^n \Rightarrow 0$ by (3.7) and (3.23). By (3.15) and (3.27), the second term satisfies

$$\sup_{0\leq t\leq T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n (g_F(w_i^n) - \alpha) \right|$$
$$\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} 1_{\{\gamma_i^n > w_i^n\}} w_i^n \sup_{1\leq i\leq E^n(T)} |g_F(w_i^n) - \alpha| \Rightarrow 0.$$

Then, the limit (3.25) follows from (3.27) and (3.28).

If $\tilde{G}^n_{23}(T) < \tilde{G}^n_{24}(T)$, there exist $1 \le i \le E^n(T)$ such that $F(w^n_i) > 1 - F(w^n_i)$, namely, $\sup_{1\le i\le E^n(T)} F(w^n_i) > 1/2$. It follows from (3.14) and (3.25) that (3.26) holds. □

Proof of Lemma 3.5. Since $w^n_i \le W^n(0)$ for $1 - Q^n(0) \le i \le 0$ (see Lemma 2.2), by (3.8) and Proposition 3.6, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1-Q^n(0)}^{0} (\gamma^n_i \wedge w^n_i) \le \tilde{Q}^n(0) W^n(0) \Rightarrow 0.$$

Since $\sup_{0\le t\le T}(t+W^n(t)) = T+W^n(T)$ (see Lemma 2.3), by (3.8) and Proposition 3.6 again,

$$\sup_{0\le t\le T} \int_t^{t+W^n(t)} \tilde{Q}^n(s)\,\mathrm{d}s \le \sup_{0\le t\le T+W^n(T)} \tilde{Q}^n(t) \sup_{0\le t\le T} W^n(t) \Rightarrow 0.$$

Hence, Lemma 2.4 implies that

$$\sup_{0\le t\le T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (\gamma^n_i \wedge w^n_i) - \int_0^t \tilde{Q}^n(s)\,\mathrm{d}s \right| \Rightarrow 0. \tag{3.29}$$

By Proposition 3.4, Lemma 3.6, and (3.13),

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(T)} 1_{\{\gamma^n_i \le w^n_i\}} w^n_i \le (\tilde{G}^n_1(T) + \tilde{G}^n_2(T)) \sup_{1\le i\le E^n(T)} w^n_i \Rightarrow 0.$$

Because

$$\sum_{i=1}^{E^n(t)} w^n_i - \sum_{i=1}^{E^n(t)} 1_{\{\gamma^n_i \le w^n_i\}} w^n_i \le \sum_{i=1}^{E^n(t)} (\gamma^n_i \wedge w^n_i) \le \sum_{i=1}^{E^n(t)} w^n_i,$$

we have

$$\sup_{0\le t\le T} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} w^n_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{E^n(t)} (\gamma^n_i \wedge w^n_i) \right| \Rightarrow 0 \quad \text{as } n \to \infty. \tag{3.30}$$

The assertion of this lemma follows from (3.29) and (3.30). □

3.3.4 Proof of Proposition 3.3

We are ready to prove Proposition 3.3. Recall that $A^n(t)$ is the number of customers who abandon the system during $(0, t]$, and $G^n(t)$ is the number of customers who have

arrived by time t but will eventually abandon the system. Clearly, $A^n(t) \leq G^n(t)$. We will establish a lower bound for A^n and prove that in diffusion scaling, this lower bound is asymptotically close to G^n.

For $t \geq 0$, define

$$\zeta^n(t) = \inf\{s \geq 0 : s + W^n(s) > t\}. \tag{3.31}$$

Because $s + W^n(s) \leq t$ for all $s < \zeta^n(t)$, by Lemmas 2.2 and 2.3, each customer arriving before time $\zeta^n(t)$ cannot be waiting in the buffer at time t. Similarly, because $s + W^n(s) > t$ for all $s > \zeta^n(t)$, a customer who arrives after time $\zeta^n(t)$ cannot be in service at t. We call $\zeta^n(t)$ the *differentiating time at* t because it is the critical epoch that separates the present waiting customers from those in service. The next proposition concerns the *differentiating time process* $\zeta^n = \{\zeta^n(t) : t \geq 0\}$.

Proposition 3.8. *Assume that (3.4) and (3.6)–(3.8) hold. Then,* $\zeta^n \in \mathbb{D}$ *is nondecreasing for each* $n \in \mathbb{N}$ *and*

$$\zeta^n \Rightarrow \zeta \quad \text{as } n \to \infty,$$

where $\zeta(t) = t$ *for* $t \geq 0$ *is the identity function on* $\mathbb{R}_+$.

Proof. Suppose that for some $0 \leq s < t$, we have $\zeta^n(t) < \zeta^n(s)$. Then, by (3.31),

$$t < u + W^n(u) \leq s \quad \text{for } \zeta^n(t) < u < \zeta^n(s),$$

leading to a contradiction. It means that ζ^n is nondecreasing.

Since $\zeta^n(t) \leq t$ and ζ^n is nondecreasing, $\zeta^n(t-)$ exists for each $t > 0$. Fix $\varepsilon > 0$ and $t \geq 0$. We have

$$\zeta^n(t) + \varepsilon + W^n(\zeta^n(t) + \varepsilon) > t + \delta \quad \text{for some } \delta > 0,$$

so that $\zeta^n(t+\delta_0) \leq \zeta^n(t+\delta) \leq \zeta^n(t) + \varepsilon$ for $0 < \delta_0 \leq \delta$. Hence, ζ^n is right-continuous at t and thus $\zeta^n \in \mathbb{D}$.

Because W^n is right-continuous,

$$\zeta^n(t) + W^n(\zeta^n(t)) \geq t \quad \text{for } t \geq 0,$$

which, along with the fact $\zeta^n(t) \leq t$, implies that

$$\sup_{0\leq t\leq T} |t - \zeta^n(t)| \leq \sup_{0\leq t\leq T} W^n(\zeta^n(t)) \leq \sup_{0\leq t\leq T} W^n(t) \quad \text{for any } T > 0.$$

It follows from Proposition 3.6 that $\zeta^n \Rightarrow \zeta$ as $n \to \infty$. □

Proof of Proposition 3.3. By (3.9), (3.12), and Propositions 3.4–3.5, we deduce that the convergence (3.11) holds. Because each customer arriving before time $\zeta^n(t)$ should have entered service or have abandoned the system by time t,

$$G^n(\zeta^n(t) - \delta) \leq A^n(t) \quad \text{for all } \delta > 0,$$

where by convention, we take $G^n(u) = 0$ for $u < 0$. Setting $\delta = n^{-1}$, we obtain both lower and upper bounds of A^n,

$$G^n(\zeta^n(t) - n^{-1}) \leq A^n(t) \leq G^n(t). \tag{3.32}$$

In diffusion scaling,

$$\begin{aligned}\tilde{G}^n(t) - \tilde{G}^n(\zeta^n(t) - n^{-1}) \leq \Big|\tilde{G}^n(t) - \alpha \int_0^t \tilde{Q}^n(s)\,\mathrm{d}s\Big| + \alpha \int_{(\zeta^n(t)-n^{-1})^+}^t \tilde{Q}^n(s)\,\mathrm{d}s \\ + \Big|\tilde{G}^n(\zeta^n(t) - n^{-1}) - \alpha \int_0^{(\zeta^n(t)-n^{-1})^+} \tilde{Q}^n(s)\,\mathrm{d}s\Big|.\end{aligned}$$

Because $\zeta^n(t) \leq t$,

$$\sup_{0\leq t\leq T} \Big|\tilde{G}^n(\zeta^n(t) - n^{-1}) - \alpha \int_0^{(\zeta^n(t)-n^{-1})^+} \tilde{Q}^n(s)\,\mathrm{d}s\Big| \leq \sup_{0\leq t\leq T} \Big|\tilde{G}^n(t) - \alpha \int_0^t \tilde{Q}^n(s)\,\mathrm{d}s\Big|.$$

It follows that

$$\begin{aligned}\sup_{0\leq t\leq T} \big\{\tilde{G}^n(t) - \tilde{G}^n(\zeta^n(t) - n^{-1})\big\} \leq 2\Big|\tilde{G}^n(t) - \alpha \int_0^t \tilde{Q}^n(s)\,\mathrm{d}s\Big| \\ + \alpha \int_{(\zeta^n(t)-n^{-1})^+}^t \tilde{Q}^n(s)\,\mathrm{d}s.\end{aligned}$$

By (3.8) and Lemma 3.8,

$$\sup_{0\le t\le T}\int_{(\zeta^n(t)-n^{-1})^+}^{t}\tilde{Q}^n(s)\,ds \le \sup_{0\le t\le T}|t-\zeta^n(t)|\sup_{0\le t\le T}\tilde{Q}^n(t)+\frac{1}{n}\sup_{0\le t\le T}\tilde{Q}^n(t)\Rightarrow 0,$$

which, together with (3.11), yields

$$\sup_{0\le t\le T}\left\{\tilde{G}^n(t)-\tilde{G}^n(\zeta^n(t)-n^{-1})\right\}\Rightarrow 0 \quad \text{as } n\to\infty. \tag{3.33}$$

Finally, the proposition follows from (3.32) and (3.33). □

3.4 Proof of Theorem 3.2

To prove Theorem 3.2, for $\ell = 1, 2$ and $i \ge 1 - N(0)$, let $r_i^{(\ell)}(t)$ be the remaining service time of the ith customer in $\Sigma^{(\ell)}$ at time t. Recall that $E(t)$ is the number of customer arrivals to both queues in $(0, t]$.

Proof of Theorem 3.2. The set of customers being served in $\Sigma^{(\ell)}$ at time $t \ge 0$ can be represented by

$$\Pi^{(\ell)}(t)=\left\{i\in\mathbb{Z}: 1-N(0)\le i\le E(t),\ r_i^{(\ell)}(t)>0,\ \sum_{k=1-N(0)}^{i}1_{\{r_k^{(\ell)}(t)>0\}}\le n\right\}. \tag{3.34}$$

Set $\xi_0 = 0$ and let $0 < \xi_1 \le \xi_2 \le \cdots$ be the service completion times in $\Sigma^{(2)}$. By (3.10), at time $\xi_0 = 0$, $r_i^{(1)}(\xi_0) \le r_i^{(2)}(\xi_0)$ for all $i \ge 1 - N(0)$.

Suppose that $r_i^{(1)}(t) \le r_i^{(2)}(t)$ for $0 \le t \le \xi_m$. For any $i \in \Pi^{(2)}(\xi_m)$, representation (3.34) implies either $i \in \Pi^{(1)}(\xi_m)$ or $r_i^{(1)}(\xi_m) = 0$. Since for $t \in (\xi_m, \xi_{m+1}]$, $r_i^{(2)}(t) = r_i^{(2)}(\xi_m) - (t - \xi_m) \ge 0$ and $r_i^{(1)}(t) = (r_i^{(1)}(\xi_m) - (t - \xi_m))^+$, then $r_i^{(1)}(t) \le r_i^{(2)}(t)$. If $i \notin \Pi^{(2)}(\xi_m)$, $r_i^{(1)}(t) \le r_i^{(2)}(t)$ also holds for $t \in (\xi_m, \xi_{m+1}]$ because $r_i^{(2)}(t) = r_i^{(2)}(\xi_m)$ and $r_i^{(1)}(t) \le r_i^{(1)}(\xi_m)$. By induction, we get $r_i^{(1)}(t) \le r_i^{(2)}(t)$ for all $t \ge 0$ and $i \ge 1 - N(0)$.

For $t \ge 0$, let

$$M(t)=\begin{cases}\max\{i\in\Pi^{(2)}(t)\} & \text{if } \Pi^{(2)}(t)\neq\varnothing,\\ E(t) & \text{if } \Pi^{(2)}(t)=\varnothing,\end{cases}$$

which is the index of the last customer being served during $(0,t]$ in $\Sigma^{(2)}$. So $Q^{(2)}(t) = E(t) - M(t)$ and

$$\sum_{i=1-N(0)}^{M(t)-1} 1_{\{r_i^{(2)}(t)>0\}} < n.$$

Since $r_i^{(1)}(t) \le r_i^{(2)}(t)$ for each $i \ge 1 - N(0)$, the above inequality leads to

$$\sum_{i=1-N(0)}^{M(t)-1} 1_{\{r_i^{(1)}(t)>0\}} < n,$$

which implies $Q^{(1)}(t) \le E(t) - M(t)$. Therefore, $Q^{(1)}(t) \le Q^{(2)}(t)$. □

3.5 On the initial condition

In this section, we present a proposition that gives a justification for imposing the initial condition (3.9) in Theorem 3.1. Let $G_Q^n(t)$ be the number of customers in the nth system who are waiting in the buffer at time t, but will eventually abandon the system. Clearly,

$$G^n(0) = G_Q^n(0).$$

Its diffusion-scaled version is given by

$$\tilde{G}_Q^n(t) = \frac{1}{\sqrt{n}} G_Q^n(t).$$

Regarding the process $\tilde{G}_Q^n = \{\tilde{G}_Q^n(t) : t \ge 0\}$, we have the following result.

Proposition 3.9. *Under the conditions of Theorem 3.1,*

$$\tilde{G}_Q^n \Rightarrow 0 \quad \text{as } n \to \infty.$$

Proof. By (3.31), any customer who is waiting in the buffer at time $t \ge 0$ must arrive at the system during $[\zeta^n(t), t]$ (the initial customers are regarded as arriving at time zero), and by Lemmas 2.2 and 2.3, he must leave the buffer by time $t + W^n(t)$. This implies

$$G_Q^n(t) \le A^n(t + W^n(t)) - A^n(\zeta^n(t)-),$$

where we set $A^n(0-) = 0$ by convention. It follows that for any $T > 0$,

$$\sup_{0\le t\le T} \tilde{G}_Q^n(t) \le \sup_{0\le t\le T} |\tilde{A}^n(t + W^n(t)) - \tilde{A}^n(\zeta^n(t))| + \sup_{0\le t\le T} |\tilde{A}^n(\zeta^n(t)) - \tilde{A}^n(\zeta^n(t)-)|.$$

It follows from Theorem 3.1 and Propositions 3.6 and 3.8 that

$$\sup_{0\le t\le T} |\tilde{A}^n(t + W^n(t)) - \tilde{A}^n(\zeta^n(t))| \Rightarrow 0 \quad \text{as } n \to \infty.$$

By Theorem 3.1 and the fact $\zeta^n(t) \le t$,

$$\sup_{0\le t\le T} |\tilde{A}^n(\zeta^n(t)) - \tilde{A}^n(\zeta^n(t)-)| \le \sup_{0\le t\le T} |\tilde{A}^n(t) - \tilde{A}^n(t-)| \Rightarrow 0 \quad \text{as } n \to \infty.$$

Hence, $\tilde{G}_Q^n \Rightarrow 0$ as $n \to \infty$. □

Assume that the queue is initially empty. Then, condition (3.9) is satisfied at time zero. Proposition 3.9 implies that under the conditions of Theorem 3.1, $\tilde{G}_Q^n(t) \Rightarrow 0$ as $n \to \infty$ for any $t \ge 0$. Thus, if we start to observe the system at any fixed time $t > 0$, the initial condition (3.9) is indeed satisfied at time t. In [37], an initial assumption similar to (3.9) is made for $G/GI/n + GI$ queues in the QED regime.

CHAPTER IV

DIFFUSION LIMITS FOR $G/Ph/n+GI$ QUEUES

In this chapter, we investigate limit processes for many-server queues that allow for customer abandonment. These queues are assumed to be operated in the QED regime and the service time distribution is restricted to be phase-type. More specifically, we study a sequence of $G/Ph/n+GI$ queues where the Ph signifies that the service times are iid following a phase-type distribution. Phase-type distributions can be used to approximate any positive-valued distribution [38].

In Theorem 4.1, we prove that in diffusion scaling, the customer-count processes and the server-allocation processes converge jointly to a multidimensional Markov process $(\tilde{N}, \tilde{Z})$ in distribution. Theorem 4.2 demonstrates that if we classify customers according to their first service phases, then at any time, the waiting customers are distributed following the distribution of a customer's first service phase. In Theorem 4.3, we prove that the diffusion-scaled customer-vector processes converge in distribution to a multidimensional diffusion process $\tilde{X}$. In Theorem 4.4, the diffusion-scaled virtual waiting time processes are proved to converge in distribution to a constant multiple of $\tilde{N}^+$, which serves as the limit of the diffusion-scaled queue length processes. Although the limit process $(\tilde{N}, \tilde{Z})$ in Theorem 4.1 is not a diffusion process in a strict sense, we still call it a diffusion limit because it is a simple transformation of the diffusion process $\tilde{X}$ in Theorem 4.3. This terminology is consistent with the usage in the conventional heavy traffic theory, where limit processes are often constrained diffusion processes [48].

We introduce diffusion processes in Section 4.1 and phase-type distributions in Section 4.2. Four theorems are stated in Section 4.3. The rest of the chapter is

dedicated to the proofs. The system equations of a $G/Ph/n+GI$ queue are derived in Section 4.4. The proofs of Theorems 4.1, 4.2, and 4.4 are presented in Sections 4.5, 4.6, and 4.7, respectively.

4.1 Diffusion processes

This chapter focuses on multidimensional diffusion processes (or their transformations) that serve as limit processes for $G/Ph/n+GI$ queues in the QED regime. Let d be positive integer and $(\Omega, \mathscr{F}, \mathbb{F}, \mathbb{P})$ be a filtered probability space with filtration $\mathbb{F} = \{\mathscr{F}_t : t \geq 0\}$. A *d-dimensional diffusion process* $X = \{X(t) : t \geq 0\}$ is defined to be a solution of the following stochastic differential equation

$$X(t) = X(0) + \int_0^t b(X(s))\,\mathrm{d}s + \int_0^t \sigma(X(s))\,\mathrm{d}B(s), \tag{4.1}$$

where the *drift coefficient* b is a function defined from $\mathbb{R}^d$ to $\mathbb{R}^d$, the *diffusion coefficient* σ is a function defined from $\mathbb{R}^d$ to $\mathbb{R}^{d\times m}$, and $B = \{B(t) : t \geq 0\}$ is an m-dimensional standard Brownian motion with respect to $\mathbb{F}$. Assume that both b and σ are Lipschitz continuous, i.e., there exists a constant $c_1 > 0$ such that

$$|b(x) - b(y)| + |\sigma(x) - \sigma(y)| \leq c_1|x - y| \quad \text{for all } x, y \in \mathbb{R}^d. \tag{4.2}$$

With this condition, the stochastic differential equation (4.1) has a unique *strong solution*, i.e., there exists a unique process X on $(\Omega, \mathscr{F}, \mathbb{F}, \mathbb{P})$ such that (a) X is adapted to $\mathbb{F}$, (b) for each sample path $\omega \in \Omega$, $X(t, \omega)$ is continuous in t, and (c) for each $t \geq 0$, the stochastic differential equation (4.1) holds almost surely. See [40] for more details.

4.2 Phase-type distributions

Let p be a d-dimensional nonnegative vector whose entries sum to one, ν be a d-dimensional vector with positive entries, and P be a $d \times d$ sub-stochastic matrix. We assume that the diagonal entries of P are zero, namely,

$$P_{jj} = 0 \quad \text{for } j = 1, \ldots, d, \tag{4.3}$$

and that P is transient, namely, $I - P$ is invertible. Consider a continuous-time Markov chain with $d+1$ states (or phases) where states $1, \ldots, d$ are transient and state $d+1$ is absorbing. For $j = 1, \ldots, d$, the Markov chain starts in state j with probability p_j. The amount of time it stays in state j is exponentially distributed with mean $1/\nu_j$. When it leaves state j, the Markov chain enters state $\ell = 1, \ldots, d$ with probability $P_{j\ell}$ or enters state $d+1$ with probability $1 - \sum_{\ell=1}^{d} P_{j\ell}$. One can check that the rate matrix (or the generator) of this Markov chain is given by

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

where $\check{F} = \mathrm{diag}(\nu)(P - I)$ is a $d \times d$ matrix and $\check{c} = -\check{F}e$ is a d-dimensional vector. Given condition (4.3), the rate matrix $\check{G}$ and the pair (ν, P) are uniquely determined from each other.

A *phase-type random variable* v 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 $\check{G}$ reaches the absorbing state $d+1$. Such a phase-type distribution is said to have d phases. If P is a zero matrix, the associated phase-type distribution is called a *hyperexponential distribution* with d phases. It is well known (see, e.g., [34]) that

$$\mathbb{P}[v \le z] = 1 - p' \exp(\check{F}z)e \quad \text{for } z \ge 0.$$

We may sample a phase-type random variable with parameters (p, ν, P) as follows. We first sample a sequence of phases $j_1, \ldots, j_L$ in $\mathcal{D} = \{1, \ldots, d\}$: Phase j_1 is sampled following the distribution p on $\mathcal{D}$ at the beginning. Assume that $j_1, \ldots, j_k \in \mathcal{D}$ have been sampled; then setting $j = j_k$, we sample a phase from $\{1, \ldots, d+1\}$ following a distribution that is determined by the jth row of P with the probability of getting phase $d+1$ being $1 - \sum_{\ell=1}^{d} P_{j\ell}$. The resulting phase is denoted by j_{k+1}. If $j_{k+1} = d+1$, terminate this procedure and set $L = k$. Otherwise, continue the sampling process. Because the matrix P is assumed to be transient, $L < \infty$ almost

surely. Let $\check{\xi}_1, \ldots, \check{\xi}_L$ be independently sampled from exponential distributions with respective rates $\nu_{j_1}, \ldots, \nu_{j_L}$. Then,

$$v = \sum_{i=1}^{L} \check{\xi}_i. \tag{4.4}$$

4.3 Limit theorems

We consider a sequence of $G/Ph/n+GI$ queues indexed by the number of servers n. These queues are assumed to be operated in the QED regime, i.e., condition (1.1) holds. Recall that $E^n = \{E^n(t) : t \geq 0\}$ is the arrival process of the nth queue. We assume that the arrival processes satisfy

$$\tilde{E}^n \Rightarrow \tilde{E} \quad \text{as } n \to \infty, \tag{4.5}$$

where

$$\tilde{E}^n(t) = \frac{1}{\sqrt{n}} \hat{E}^n(t), \quad \hat{E}^n(t) = E^n(t) - \lambda^n t, \tag{4.6}$$

and $\tilde{E}$ is a driftless Brownian motion. Condition (4.5) holds, for example, when $E^n(t) = E_0(\lambda^n t)$ with E_0 being a renewal process. We assume that all these queues in the sequence have the same phase-type service time distribution that has parameters (p, ν, P). By (4.4), each customer's service time can be decomposed into a number of phases. When a customer is in service, he must be in one of the d service phases. Let $Z_j^n(t)$ denote the number of customers in phase j service in the nth queue at time t. The service times in phase j are exponentially distributed with mean $1/\nu_j$. We use $Z^n(t)$ to denote the corresponding d-dimensional vector and $Z^n = \{Z^n(t) : t \geq 0\}$ is called the *server-allocation process*. Let $N^n(t)$ denote the number of customers in the nth system at time t. We call $N^n = \{N^n(t) : t \geq 0\}$ the *customer-count process*. Setting

$$\hat{N}^n(t) = N^n(t) - n, \tag{4.7}$$

one can check that at time t, the queue length of the nth system satisfies

$$Q^n(t) = \hat{N}^n(t)^+$$

and $\hat{N}^n(t)^-$ is the number of idle servers, namely,

$$\hat{N}^n(t)^- = n - e'Z^n(t). \tag{4.8}$$

The processes $\hat{N}^n$ and Z^n describe the "state" of the system as time evolves. Hereafter, they are called the *state processes* for the nth system. A diffusion-scaled version of the customer-count process is defined by

$$\tilde{N}^n(t) = \frac{1}{\sqrt{n}}\hat{N}^n(t).$$

The customers in service are distributed among the d phases following a distribution θ, given by

$$\theta = \mu R^{-1} p, \tag{4.9}$$

$$R = (I - P')\,\mathrm{diag}(\nu). \tag{4.10}$$

One can check that $\sum_{j=1}^{d} \theta_j = 1$ and θ_j is interpreted to be the fraction of phase j work load on the n servers. This suggests the following centering for the server-allocation process

$$\hat{Z}^n(t) = Z^n(t) - n\theta.$$

We define the corresponding diffusion-scaled process by

$$\tilde{Z}^n(t) = \frac{1}{\sqrt{n}}\hat{Z}^n(t).$$

We assume that the customers who arrive after time zero have iid patience times that satisfy conditions (3.4) and (3.5), i.e.,

$$F(0) = 0 \quad \text{and} \quad \alpha = \lim_{t \downarrow 0} t^{-1} F\,(t) < \infty,$$

and that there exists a $(d+1)$-dimensional random vector $(\tilde{N}(0), \tilde{Z}(0))$ such that

$$(\tilde{N}^n(0), \tilde{Z}^n(0)) \Rightarrow (\tilde{N}(0), \tilde{Z}(0)) \quad \text{as } n \to \infty, \tag{4.11}$$

where $\tilde{N}(0)$ is a random variable and $\tilde{Z}(0)$ is a d-dimensional random vector. Both $\tilde{N}(0)$ and $\tilde{Z}(0)$ are assumed to be defined on a probability space that is rich enough so that stochastic processes $\tilde{E}, \tilde{\Phi}^0, \ldots, \tilde{\Phi}^d$, and $\tilde{S}$ defined on this space are all independent of $(\tilde{N}(0), \tilde{Z}(0))$. Here, $\tilde{E}$ is a one-dimensional driftless Brownian motion; $\tilde{\Phi}^0, \ldots, \tilde{\Phi}^d$, and $\tilde{S}$ are d-dimensional driftless Brownian motions. These Brownian motions, possibly degenerate, are mutually independent and start from zero. The variance of $\tilde{E}$ is μc_a^2 for some constant $c_a^2 \geq 0$, and the covariance matrices of $\tilde{\Phi}^0, \ldots, \tilde{\Phi}^d$ and $\tilde{S}$ are $H^0, \ldots, H^d$, and $\mathrm{diag}(\nu)$, respectively, where

$$H^0_{k\ell} = \begin{cases} p_k(1-p_\ell) & \text{if } k = \ell \\ -p_k p_\ell & \text{otherwise} \end{cases} \quad \text{and} \quad H^j_{k\ell} = \begin{cases} P_{jk}(1-P_{j\ell}) & \text{if } k = \ell \\ -P_{jk}P_{j\ell} & \text{otherwise} \end{cases} \tag{4.12}$$

for $j = 1, \ldots, d$.

To state the main theorems of this chapter, let

$$\tilde{U}(t) = \tilde{N}(0) + \tilde{E}(t) - \mu\beta t + e'\tilde{M}(t), \tag{4.13}$$

$$\tilde{V}(t) = (I - pe')\tilde{Z}(0) + \tilde{\Phi}^0(\mu t) + (I - pe')\tilde{M}(t), \tag{4.14}$$

where

$$\tilde{M}(t) = \sum_{j=1}^{d} \tilde{\Phi}^j(\nu_j\theta_j t) - (I - P')\tilde{S}(\theta t).$$

The process $(\tilde{U}, \tilde{V})$ is a $(d+1)$-dimensional Brownian motion. It is degenerate because $e'\tilde{V}(t) = 0$. Before we state the first theorem of this chapter, we present the following lemma, which is a corollary of Lemma A.1 in Appendix A.

Lemma 4.1. *Let p be a d-dimensional vector that is the distribution of the initial phases of the phase-type service times, R be the $d \times d$ matrix defined by (4.10), and $\alpha \geq 0$ is the patience time density at zero defined by (3.5). (a) For each $(u, v) \in \mathbb{D}^{d+1}$ with $u(t) \in \mathbb{R}$ and $v(t) \in \mathbb{R}^d$, there exists a unique $(x, z) \in \mathbb{D}^{d+1}$ with $x(t) \in \mathbb{R}$ and*

$z(t) \in \mathbb{R}^d$, such that

$$x(t) = u(t) - \alpha \int_0^t x(s)^+ \, \mathrm{d}s - e'R \int_0^t z(s) \, \mathrm{d}s, \tag{4.15}$$

$$z(t) = v(t) - px(t)^- - (I - pe')R \int_0^t z(s) \, \mathrm{d}s. \tag{4.16}$$

(b) For each $(u, v) \in \mathbb{D}^{d+1}$, define $\Phi(u, v) = (x, z) \in \mathbb{D}^{d+1}$, where (x, z) satisfies (4.15) and (4.16). The map Φ is well defined and is continuous when both the domain and the range $\mathbb{D}^{d+1}$ are endowed with the Skorohod J_1-topology. (c) The map Φ is Lipschitz continuous in the sense that for any $T > 0$, there exists a constant $c_T > 0$ such that

$$\sup_{0 \le t \le T} |\Phi(y)(t) - \Phi(\check{y})(t)| \le c_T \sup_{0 \le t \le T} |y(t) - \check{y}(t)| \quad \text{for any } y, \check{y} \in \mathbb{D}^{d+1}.$$

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

$$\Phi(ay) = a\Phi(y) \quad \text{for each } a > 0 \text{ and each } y \in \mathbb{D}^{d+1}.$$

Recall that $G^n(0)$ is the number of customers who are waiting in the buffer at time zero but will eventually abandon the system. To state Theorem 4.1, we assume that condition (3.9) holds, i.e.,

$$\tilde{G}^n(0) \Rightarrow 0 \quad \text{as } n \to \infty.$$

Theorem 4.1. *Consider a sequence of $G/Ph/n+GI$ queues in the QED regime, i.e., condition (1.1) holds. Assume that conditions (3.4), (3.5), (3.9), (4.5), and (4.11) also hold. Then,*

$$(\tilde{N}^n, \tilde{Z}^n) \Rightarrow (\tilde{N}, \tilde{Z}) \quad \text{as } n \to \infty,$$

where

$$(\tilde{N}, \tilde{Z}) = \Phi(\tilde{U}, \tilde{V}). \tag{4.17}$$

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 $j = 1, \ldots, d$, we use $\mathcal{Q}_j^n(t)$ 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 $\mathcal{Q}_j^n(t) = 0$. We use $Y_j^n(t)$ to denote the number of phase j customers in the system at time t, i.e.,

$$Y_j^n(t) = \mathcal{Q}_j^n(t) + Z_j^n(t).$$

Let $\mathcal{Q}^n(t)$ and $Y^n(t)$ denote the corresponding d-dimensional vectors. Set

$$\tilde{\mathcal{Q}}^n(t) = \frac{1}{\sqrt{n}}\mathcal{Q}^n(t), \quad \tilde{Y}^n(t) = \frac{1}{\sqrt{n}}\hat{Y}^n(t), \quad \hat{Y}^n(t) = Y^n(t) - n\theta.$$

Clearly,

$$\tilde{Y}^n(t) = \tilde{\mathcal{Q}}^n(t) + \tilde{Z}^n(t) \quad \text{and} \quad \tilde{N}^n(t) = e'\tilde{Y}^n(t).$$

The following theorem says that for the sequence of queues in the QED regime, the waiting customers are distributed among the d phases following distribution p. It is known as a state space collapse result.

Theorem 4.2. *Under the conditions of Theorem 4.1, for any $T > 0$,*

$$\sup_{0\le t\le T} \left|\tilde{\mathcal{Q}}^n(t) - p\tilde{N}^n(t)^+\right| \Rightarrow 0 \quad \textit{as } n \to \infty.$$

The following theorem is a corollary to Theorems 4.1 and 4.2. When there is no customer abandonment and the arrival process is renewal, it is identical to Theorem 2.3 of [43].

Theorem 4.3. *Under the conditions of Theorem 4.1,*

$$(\tilde{N}^n, \tilde{Y}^n, \tilde{Z}^n) \Rightarrow (\tilde{N}, \tilde{X}, \tilde{Z}) \quad \textit{as } n \to \infty,$$

where $(\tilde{N}, \tilde{Z})$ is defined in (4.17) and

$$\tilde{X}(t) = p\tilde{N}(t)^+ + \tilde{Z}(t).$$

The process $\tilde{X}$ satisfies

$$\tilde{X}(t) = \tilde{X}(0) - \beta\mu pt + \tilde{\Phi}^0(\mu t) + p\tilde{E}(t) + \tilde{M}(t) \\ - R\int_0^t \tilde{X}(s)\,\mathrm{d}s + (R-\alpha I)p\int_0^t (e'\tilde{X}(s))^+\,\mathrm{d}s. \quad (4.18)$$

The process $\tilde{X}$ in Theorem 4.3 is a diffusion process. Hence, it is a continuous Markov process. The drift coefficient of $\tilde{X}$ is

$$b(x) = -\beta\mu p - R(x - p(e'x)^+) - p\alpha(e'x)^+ \quad \text{for } x \in \mathbb{R}^d \quad (4.19)$$

and the diffusion coefficient σ is a $d \times d$ constant matrix satisfying

$$\Sigma(x) = \sigma(x)\sigma'(x) = \mu(c_a^2 pp' + H^0) + \sum_{j=1}^{d} \nu_j\theta_j H^j + (I - P')\,\mathrm{diag}(\nu)\,\mathrm{diag}(\theta)(I - P).$$

The drift coefficient b in (4.19) is a piecewise linear function of x. Both b and σ are Lipschitz continuous. Therefore, a strong solution to (4.18) exists and it is known as a d-dimensional piecewise Ornstein–Uhlenbeck (OU) process.

The map Φ in (4.17) defines $(\tilde{N}, \tilde{Z})$ as a $(d+1)$-dimensional continuous process, which is degenerate because it lives on a d-dimensional manifold. From the d-dimensional process $\tilde{X}$, one can recover the $(d+1)$-dimensional process $(\tilde{N}, \tilde{Z})$ via

$$\tilde{N}(t) = e'\tilde{X}(t) \quad \text{and} \quad \tilde{Z}(t) = \tilde{X}(t) - p\tilde{N}(t)^+. \quad (4.20)$$

Therefore, $(\tilde{N}, \tilde{Z})$ is a continuous Markov process. However, the process $(\tilde{N}, \tilde{Z})$ is not a diffusion process by the common definition in Section 4.1, because the function x^+ in (4.20) is not differentiable at zero. As a consequence, the drift and the diffusion coefficients of $(\tilde{N}, \tilde{Z})$ are not well defined at zero. A similar observation was made by [57, Remark 2.2]: The limit process there is not a diffusion process either, but a simple transformation of that limit process is a diffusion process.

Our next theorem is concerned with the virtual waiting time process W^n. When there is no customer abandonment and the arrival processes are renewal, the theorem is implied by Corollary 2.3 and Remark 2.6 of [43].

Theorem 4.4. *Under the conditions of Theorem 4.1,*

$$\sqrt{n}W^n \Rightarrow \frac{\tilde{N}^+}{\mu} \quad as\ n \to \infty.$$

The basic tools in the proofs of the above theorems include the standard FCLT, the continuous-mapping theorem, and the random-time-change theorem. First, the system equations for a $G/Ph/n + GI$ queue are derived to construct a continuous map. As the main result of this chapter, Theorem 4.1 is first proved for $G/Ph/n$ queues without customer abandonment. This helps us obtain the stochastic boundedness of the diffusion-scaled queue length processes. Then, we prove the theorem for the $G/Ph/n + GI$ model by exploiting the asymptotic relationship in (1.2) and the comparison result in Theorem 3.2. Theorem 4.3 is a corollary to Theorems 4.1 and 4.2. The proofs of Theorems 4.2 and 4.4 also rely extensively on the convergence results proved in Chapter 3.

4.4 System equations

The $G/Ph/n + GI$ queue is driven by several primitive processes. For $j = 1, \ldots, d$, let $S_j = \{S_j(t) : t \geq 0\}$ be a Poisson process with rate ν_j and $\phi^j = \{\phi^j(i) : i \in \mathbb{N}\}$ be a sequence of iid d-dimensional random vectors such that $\phi^j(i)$ takes e^ℓ with probability $P_{j\ell}$ and takes a d-dimensional zero vector with probability $1 - \sum_{\ell=1}^d P_{j\ell}$. Similarly, let $\phi^0 = \{\phi^0(i) : i \in \mathbb{N}\}$ be a sequence of iid d-dimensional random vectors such that $\phi^0(i)$ takes e^ℓ with probability p_ℓ. For $j = 0, \ldots, d$, define the routing process $\Phi^j = \{\Phi^j(k) : k \in \mathbb{N}\}$ by

$$\Phi^j(k) = \sum_{i=1}^{k} \phi^j(i).$$

We assume that $N^n(0), E^n, S_1, \ldots, S_d, \Phi^0, \ldots, \Phi^d$ are mutually independent.

For $j = 1, \ldots, d$, let $T_j^n(t)$ be the cumulative amount of service effort received by customers in phase j service by time t, $B^n(t)$ be the cumulative number of customers

who have entered service in $(0,t]$, and $D^n(t)$ be the cumulative number of customers who have completed service in $(0,t]$. Clearly,

$$T_j^n(t) = \int_0^t Z_j^n(s)\,\mathrm{d}s.$$

Thus, $S_j(T_j^n(t))$ is equal in distribution to the cumulative number of phase j service completions by time t. (Please refer to Section 4.1 of [11] on a perturbed system and [53] for a more general treatment.) Recall that $A^n(t)$ is the cumulative number of customers who have abandoned the system by time t. One can check that the processes N^n and Z^n satisfy the following dynamical equations,

$$Z^n(t) = Z^n(0) + \Phi^0(B^n(t)) + \sum_{j=1}^{d} \Phi^j(S_j(T_j^n(t))) - S(T^n(t)), \tag{4.21}$$

$$N^n(t) = N^n(0) + E^n(t) - D^n(t) - A^n(t), \tag{4.22}$$

$$D^n(t) = -e'\bigg(\sum_{j=1}^{d} \Phi^j(S_j(T_j^n(t))) - S(T^n(t))\bigg), \tag{4.23}$$

where

$$S(T^n(t)) = (S_1(T_1^n(t)), \ldots, S_d(T_d^n(t)))'.$$

We define the following centered processes

$$\hat{S}(t) = S(t) - \nu t, \quad \hat{\Phi}^0(k) = \sum_{i=1}^{k} \left(\phi^0(i) - p\right), \quad \hat{\Phi}^j(k) = \sum_{i=1}^{k} \left(\phi^j(i) - p^j\right)$$

for $j = 1, \ldots, d$ and $k \in \mathbb{N}$, where p^j is the jth column of P'. Setting

$$M^n(t) = \sum_{j=1}^{d} \hat{\Phi}^j\left(S_j(T_j^n(t))\right) - (I - P')\hat{S}\left(T^n(t)\right), \tag{4.24}$$

one has

$$\sum_{j=1}^{d} \Phi^j(S_j(T_j^n(t))) - S(T^n(t)) = M^n(t) - R\int_0^t Z^n(s)\,\mathrm{d}s,$$

where R is defined in (4.10). By (4.21) and (4.23),

$$e'Z^n(t) = e'Z^n(0) + B^n(t) - D^n(t), \tag{4.25}$$

$$D^n(t) = -e'M^n(t) + e'R\int_0^t Z^n(s)\,\mathrm{d}s. \tag{4.26}$$

It follows from (4.7), (4.8), and (4.21)–(4.26) that

$$\begin{aligned}\hat{N}^n(t) &= \hat{N}^n(0) + \hat{E}^n(t) + \lambda^n t + e'M^n(t) - e'R\int_0^t Z^n(s)\,\mathrm{d}s - A^n(t),\\ Z^n(t) &= Z^n(0) + p\hat{N}^n(0)^- + \hat{\Phi}^0(B^n(t)) - p\hat{N}^n(t)^-\\ &\quad + (I-pe')M^n(t) - (I-pe')R\int_0^t Z^n(s)\,\mathrm{d}s.\end{aligned}$$

Recall that $\hat{Z}^n(t) = Z^n(t) - n\theta$. We then have

$$\begin{aligned}\hat{N}^n(t) &= \hat{N}^n(0) + \hat{E}^n(t) + (\lambda^n - n\mu)t + e'M^n(t) - e'R\int_0^t \hat{Z}^n(s)\,\mathrm{d}s - A^n(t),\\ \hat{Z}^n(t) &= (I-pe')\hat{Z}^n(0) + \hat{\Phi}^0(B^n(t)) - p\hat{N}^n(t)^-\\ &\quad + (I-pe')M^n(t) - (I-pe')R\int_0^t \hat{Z}^n(s)\,\mathrm{d}s,\end{aligned}$$

where we have used (4.9) and (4.10) in the derivation. Setting

$$U^n(t) = \hat{N}^n(0) + \hat{E}^n(t) + (\lambda^n - n\mu)t + e'M^n(t) - A^n(t) + \alpha\int_0^t \hat{N}^n(s)^+\,\mathrm{d}s, \quad (4.27)$$

$$V^n(t) = (I-pe')\hat{Z}^n(0) + \hat{\Phi}^0(B^n(t)) + (I-pe')M^n(t), \quad (4.28)$$

we finally have

$$\begin{aligned}\hat{N}^n(t) &= U^n(t) - \alpha\int_0^t \hat{N}^n(s)^+\,\mathrm{d}s - e'R\int_0^t \hat{Z}^n(s)\,\mathrm{d}s,\\ \hat{Z}^n(t) &= V^n(t) - p\hat{N}^n(t)^- - (I-pe')R\int_0^t \hat{Z}^n(s)\,\mathrm{d}s.\end{aligned}$$

By Lemma 4.1, we have obtained the following representation for the state processes

$$(\hat{N}^n, \hat{Z}^n) = \Phi(U^n, V^n). \quad (4.29)$$

4.5 Proof of Theorem 4.1

This section provides the proof of Theorem 4.1. We first prove the theorem for a sequence of $G/Ph/n$ queues without abandonment using the continuous map Φ. The fluid limits of several performance processes are established in the proof, allowing us to apply the random-time-change theorem. The resulting limit process implies

the stochastic boundedness of the diffusion-scaled queue length processes in these $G/Ph/n$ queues. By the comparison result in Theorem 3.2, the queue length processes in the corresponding queues with abandonment are stochastically bounded in diffusion scaling automatically. This enables the replacement of the abandonment process by the queue length integral in the limit process, according to the asymptotic relationship in (1.2). Then, we set up new fluid limits and apply the continuous map Φ again to finish the proof for the $G/Ph/n+GI$ model.

4.5.1 Proof for $G/Ph/n$ queues

Let us first focus on a sequence of $G/Ph/n$ queues without customer abandonment. In this case, the abandonment process $A^n = 0$ and $\alpha = 0$ in (3.5).

We define the fluid-scaled processes $\bar{B}^n$, $\bar{D}^n$, $\bar{E}^n$, $\bar{T}^n$, $\bar{K}^n$, and $\bar{Z}^n$ by

$$\begin{aligned}
\bar{B}^n(t) &= \frac{1}{n}B^n(t), \quad \bar{D}^n(t) = \frac{1}{n}D^n(t), \quad \bar{E}^n(t) = \frac{1}{n}E^n(t),\\
\bar{T}^n(t) &= \frac{1}{n}T^n(t), \quad \bar{K}^n(t) = \frac{1}{n}\hat{N}^n(t), \quad \bar{Z}^n(t) = \frac{1}{n}Z^n(t).
\end{aligned}$$

Lemma 4.2. *Consider a sequence of $G/Ph/n$ queues that satisfies (1.1) and (4.5). Assume that (4.11) holds. Then,*

$$(\bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{T}^n, \bar{K}^n, \bar{Z}^n) \Rightarrow (\bar{B}, \bar{D}, \bar{E}, \bar{T}, \bar{K}, \bar{Z}) \quad \text{as } n \to \infty,$$

where $\bar{B}(t) = \bar{D}(t) = \bar{E}(t) = \mu t$, $\bar{T}(t) = \theta t$, $\bar{K}(t) = 0$, and $\bar{Z}(t) = \theta$ for $t \geq 0$.

Proof. It follows from (1.1) and (4.5) that

$$\bar{E}^n \Rightarrow \bar{E} \quad \text{as } n \to \infty. \tag{4.30}$$

Let

$$\begin{aligned}
\bar{M}^n(t) &= \frac{1}{n}M^n(t), \quad \bar{U}^n(t) = \frac{1}{n}U^n(t), \quad \bar{V}^n(t) = \frac{1}{n}V^n(t),\\
\bar{S}^n(t) &= \frac{1}{n}S^n(nt), \quad \bar{L}^n(t) = \frac{1}{n}\hat{Z}^n(t).
\end{aligned}$$

We would next show that

$$(\bar{M}^n, \bar{U}^n, \bar{V}^n) \Rightarrow (0,0,0) \quad \text{as } n \to \infty. \tag{4.31}$$

By the FLLN, as $n \to \infty$,

$$\bar{S}^n \Rightarrow \bar{S}, \quad \frac{1}{n} \sup_{0 \le t \le T} |\hat{S}(nt)| \Rightarrow 0, \quad \frac{1}{n} \sup_{0 \le t \le T} |\hat{\Phi}^j(\lfloor nt \rfloor)| \Rightarrow 0 \tag{4.32}$$

where $\bar{S}(t) = \nu t$ and $j = 0, \ldots, d$. This, along with (4.24) and the fact $\bar{T}_j^n(t) \le t$ for $j = 1, \ldots, d$, implies that

$$\sup_{0 \le t \le T} |\bar{M}^n(t)| \Rightarrow 0 \quad \text{as } n \to \infty. \tag{4.33}$$

Note that $\bar{B}^n(t) \le \bar{K}^n(0)^+ + \bar{E}^n(t)$. By (4.11) and (4.30), the sequence of processes $\{\bar{B}^n : n \in \mathbb{N}\}$ must be stochastically bounded, i.e., for each $T > 0$,

$$\lim_{a \to \infty} \limsup_{n \to \infty} \mathbb{P}\Big[\sup_{0 \le t \le T} \bar{B}^n(t) > a \Big] = 0.$$

This, along with (4.32), leads to

$$\sup_{0 \le t \le T} \frac{1}{n} \hat{\Phi}^0 \left(B^n(t) \right) \Rightarrow 0 \quad \text{as } n \to \infty. \tag{4.34}$$

Condition (4.11) implies that $\bar{L}^n(0) \Rightarrow 0$ and $\bar{K}^n(0) \Rightarrow 0$ as $n \to \infty$, which, together with (4.33) and (4.34), leads to $\bar{V}^n \Rightarrow 0$ as $n \to \infty$. Similarly, using the fact $A^n = 0$, one can argue that $\bar{U}^n \Rightarrow 0$ as $n \to \infty$.

By (4.29) and the positively homogeneous property of Φ, $(\bar{K}^n, \bar{L}^n) = \Phi(\bar{U}^n, \bar{V}^n)$. By (4.31) and the continuity of the map Φ, we get

$$(\bar{K}^n, \bar{L}^n) \Rightarrow (0,0) \quad \text{as } n \to \infty.$$

Because $\bar{Z}^n(t) = \bar{L}^n(t) + \theta$, one has $\bar{Z}^n \Rightarrow \bar{Z}$ as $n \to \infty$, from which it follows that $\bar{T}^n \Rightarrow \bar{T}$ as $n \to \infty$. By (4.26),

$$\bar{D}^n(t) = -e' \bar{M}^n(t) + e' R \int_0^t \bar{Z}^n(s) \, \mathrm{d}s.$$

Since $e'R\int_0^t \bar{Z}(s)\,\mathrm{d}s = \mu t$, then $\bar{D}^n \Rightarrow \bar{D}$ as $n \to \infty$ by the continuous-mapping theorem. The convergence of $\bar{D}^n$ and (4.25) imply that $\bar{B}^n \Rightarrow \bar{B}$ as $n \to \infty$, where $\bar{B}$ satisfies

$$e'\bar{Z}(t) = e'\bar{Z}(0) + \bar{B}(t) - \bar{D}(t).$$

Since $\bar{Z}(t) = \bar{Z}(0) = \theta$, we conclude that $\bar{B}(t) = \mu t$. □

Define the diffusion-scaled processes $\tilde{\Phi}^{0,n}, \ldots, \tilde{\Phi}^{d,n}$, and $\tilde{S}^n$ via

$$\tilde{\Phi}^{j,n}(t) = \frac{1}{\sqrt{n}}\hat{\Phi}^j(\lfloor nt \rfloor) \quad \text{and} \quad \tilde{S}^n(t) = \frac{1}{\sqrt{n}}\hat{S}(nt).$$

By the FCLT, one has

$$(\tilde{\Phi}^{0,n}, \ldots, \tilde{\Phi}^{d,n}, \tilde{S}^n) \Rightarrow (\tilde{\Phi}^0, \ldots, \tilde{\Phi}^d, \tilde{S}) \quad \text{as } n \to \infty,$$

where $\tilde{\Phi}^0, \ldots, \tilde{\Phi}^d$, and $\tilde{S}$ are d-dimensional driftless Brownian motions. As mentioned previously, for $j = 0, \ldots, d$ the covariance matrix for $\tilde{\Phi}^j$ is H^j given by (4.12) and the covariance matrix for $\tilde{S}$ is $\mathrm{diag}(\nu)$. By assumption (4.5) for the arrival process E^n, the initial condition (4.11), and the independence assumption of primitive processes, one has

$$(\tilde{X}^n(0), \tilde{Z}^n(0), \tilde{E}^n, \tilde{\Phi}^{0,n}, \ldots, \tilde{\Phi}^{d,n}, \tilde{S}^n) \Rightarrow (\tilde{X}(0), \tilde{Z}(0), \tilde{E}, \tilde{G}, \tilde{\Phi}^0, \ldots, \tilde{\Phi}^d, \tilde{S}) \qquad (4.35)$$

as $n \to \infty$. The components of $(\tilde{E}, \tilde{G}, \tilde{S}, \tilde{\Phi}^0, \ldots, \tilde{\Phi}^K)$ are mutually independent and they are independent of $(\tilde{X}(0), \tilde{Z}(0))$.

We further define

$$\tilde{U}^n(t) = \frac{1}{\sqrt{n}}U^n(t) \quad \text{and} \quad \tilde{V}^n(t) = \frac{1}{\sqrt{n}}V^n(t).$$

These processes have the following convergence result.

Lemma 4.3. *Consider a sequence of $G/Ph/n$ queues that satisfies (1.1) and (4.5). Assume that (4.11) holds. Then,*

$$(\tilde{U}^n, \tilde{V}^n) \Rightarrow (\tilde{U}, \tilde{V}) \quad \textit{as } n \to \infty,$$

where $(\tilde{U}, \tilde{V})$ is the $(d+1)$-dimensional Brownian motion defined by (4.13) and (4.14).

Proof. Because $A^n = 0$ and $\alpha = 0$, it follows from (4.27) and (4.28) that

$$\tilde{U}^n(t) = \tilde{N}^n(0) + \tilde{E}^n(t) + \sqrt{n}\Big(\frac{1}{n}\lambda^n - \mu\Big) + e'\tilde{M}^n(t), \tag{4.36}$$

$$\tilde{V}^n(t) = (I - pe')\tilde{Z}^n(0) + \tilde{\Phi}^{0,n}(\bar{B}^n(t)) + (I - pe')\tilde{M}^n(t), \tag{4.37}$$

where

$$\tilde{M}^n(t) = \frac{1}{\sqrt{n}}M^n(t) = \sum_{j=1}^{d} \tilde{\Phi}^{j,n}(\bar{S}_j^n(\bar{T}_j^n(t))) - (I - P')\tilde{S}^n(\bar{T}^n(t)).$$

Note that $\bar{S}^n \Rightarrow \bar{S}$ as $n \to \infty$ where $\bar{S}(t) = \nu t$. This lemma follows from (4.35), Lemma 4.2, the continuous-mapping theorem, and the random-time-change theorem. □

Proof of Theorem 4.1 for $G/Ph/n$ queues. It follows from the state-process representation (4.29) and the positively homogeneous property of the map Φ that

$$(\tilde{N}^n, \tilde{Z}^n) = \Phi(\tilde{U}^n, \tilde{V}^n).$$

The theorem now follows from Lemma 4.3 and the continuous-mapping theorem. □

4.5.2 Proof for $G/Ph/n + GI$ queues

Now we turn to the $G/Ph/n + GI$ queues with customer abandonment. First, we show that the asymptotic relationship presented in Chapter 3 holds in the sequence of queues.

Lemma 4.4. *Under the conditions of Theorem 4.1, for any $T > 0$,*

$$\sup_{0\le t\le T}\Big|\tilde{A}^n(t) - \alpha\int_0^t \tilde{N}^n(s)^+\,\mathrm{d}s\Big| \Rightarrow 0 \quad \textit{as } n \to \infty.$$

Proof. In order to apply Theorem 3.1, we need only verify that the sequence of diffusion-scaled queue length processes is stochastically bounded, i.e., for any $T > 0$,

$$\lim_{a\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big[\sup_{0\le t\le T}\tilde{N}^n(t)^+ > a\Big] = 0. \tag{4.38}$$

We have proved in Section 4.5.1 that Theorem 4.1 holds for the sequence of $G/Ph/n$ queues. As a consequence, the sequence of diffusion-scaled queue length processes in the $G/Ph/n$ queues is stochastically bounded. Then, (4.38) follows from the comparison result in Theorem 3.2. □

The next lemma is an extension of the fluid limits in Lemma 4.2 that allows for customer abandonment.

Lemma 4.5. *Under the conditions of Theorem 4.1,*

$$(\bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{T}^n, \bar{K}^n, \bar{Z}^n) \Rightarrow (\bar{B}, \bar{D}, \bar{E}, \bar{T}, \bar{K}, \bar{Z}) \quad \textit{as } n \to \infty.$$

The proof of Lemma 4.5 mostly follows the proof of Lemma 4.2. The only modification is that Lemma 4.4 is used in proving $\bar{U}^n \Rightarrow 0$ as $n \to \infty$.

Proof of Theorem 4.1. Using the representation (4.29), one has

$$(\tilde{X}^n, \tilde{Z}^n) = \Phi(\tilde{U}^n, \tilde{V}^n),$$

where

$$\tilde{U}^n(t) = \tilde{N}^n(0) + \tilde{E}^n(t) + \sqrt{n}\Big(\frac{1}{n}\lambda^n - \mu\Big) + e'\tilde{M}^n(t) - \tilde{A}^n(t) + \alpha \int_0^t \tilde{N}^n(s)^+ \,\mathrm{d}s \quad (4.39)$$

and $\tilde{V}^n$ is given by (4.37). By Lemma 4.1, the map Φ is continuous. Thus, to prove the theorem, it suffices to prove that

$$(\tilde{U}^n, \tilde{V}^n) \Rightarrow (\tilde{U}, \tilde{V}) \quad \text{as } n \to \infty, \quad (4.40)$$

where $(\tilde{U}, \tilde{V})$ is the $(d+1)$-dimensional Brownian motion defined by (4.13) and (4.14). This convergence follows from the proof of Lemma 4.3 with the following two modifications. First, compared with (4.36), the extra terms of $\tilde{U}^n$ in (4.39) satisfies

$$\sup_{0 \le t \le T} \Big| \tilde{A}^n(t) - \alpha \int_0^t \tilde{N}^n(s)^+ \,\mathrm{d}s \Big| \Rightarrow 0 \quad \text{as } n \to \infty$$

by Lemma 4.4. Second, we use Lemma 4.5 instead of Lemma 4.2 in order to apply the random-time-change theorem to finish the proof of (4.40). □

4.6 Proof of Theorem 4.2

This section is devoted to proving Theorem 4.2. We first present two lemmas. The first lemma is an immediate result from Theorem 4.1 and Lemma 4.4.

Lemma 4.6. *Under the conditions of Theorem 4.1,*

$$\tilde{A}^n \Rightarrow \tilde{A} \quad \textit{as } n \to \infty,$$

where

$$\tilde{A}(t) = \alpha \int_0^t \tilde{N}(s)^+ \, \mathrm{d}s \quad \textit{for } t \geq 0.$$

Recall that ζ^n, defined in (3.31), is the differentiating time process of the nth queue. The second lemma concerns the number of customers who abandon the system during $(\zeta^n(t), t]$.

Lemma 4.7. *Under the conditions of Theorem 4.1,*

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t)) \right| \Rightarrow 0 \quad \textit{as } n \to \infty.$$

Proof. By Proposition 3.8, the differentiating time processes satisfy $\zeta^n \Rightarrow \zeta$ as $n \to \infty$ where $\zeta(t) = t$. Then, the present lemma follows from Lemma 4.6, Theorem 3.9 in [4], and the random-time-change theorem. □

Proof of Theorem 4.2. Let $T > 0$ and $t \in [0, T]$ be fixed. Let $\{\psi^0(i) : i \in \mathbb{N}\}$ be a sequence of iid d-dimensional random vectors indicating the first service phases of all customers, including those waiting in the buffer at time zero. For $j = 1, \ldots, d$, we take $\psi^0(i) = e^j$ with probability p_j. Write

$$\Psi^0(k) = \sum_{i=1}^{k} \psi^0(i) \quad \text{and} \quad \hat{\Psi}^0(k) = \Psi^0(k) - pk.$$

By the FCLT,

$$\tilde{\Psi}^{0,n} \Rightarrow \tilde{\Psi}^0 \quad \text{as } n \to \infty, \tag{4.41}$$

where

$$\tilde{\Psi}^{0,n}(t) = \frac{1}{\sqrt{n}}\hat{\Psi}^0(\lfloor nt \rfloor)$$

and $\tilde{\Psi}^0$ is a d-dimensional Brownian motion.

We first consider the case $\zeta^n(t) = 0$. Note that $\hat{N}^n(t)^+ = Q^n(t)$ is the scaled queue length at time t. Hence,

$$\hat{N}^n(t)^+ = \hat{N}^n(0)^+ + E^n(t) - B^n(t) - (A^n(t) - A^n(\zeta^n(t))), \tag{4.42}$$

where by assumption $A^n(\zeta^n(t)) = A^n(0) = 0$. Then, for $j = 1, \ldots, d$, we have

$$Q_j^n(t) \le \Psi_j^0(\hat{N}^n(0)^+ + E^n(t)) - \Psi_j^0(B^n(t)) \tag{4.43}$$

and

$$Q_j^n(t) \ge \Psi_j^0(\hat{N}^n(0)^+ + E^n(t)) - \Psi_j^0(B^n(t)) - (A^n(t) - A^n(\zeta^n(t))). \tag{4.44}$$

By (4.42)–(4.44), we obtain

$$\Lambda_j^n(t) \le Q_j^n(t) - p\hat{N}^n(t)^+ \le \Pi_j^n(t) \tag{4.45}$$

where

$$\Lambda_j^n(t) = \hat{\Psi}_j^0(\hat{N}^n(0)^+ + E^n(t)) - \hat{\Psi}_j^0(B^n(t)) - (A^n(t) - A^n(\zeta^n(t))),$$
$$\Pi_j^n(t) = \hat{\Psi}_j^0(\hat{N}^n(0)^+ + E^n(t)) - \hat{\Psi}_j^0(B^n(t)) + (A^n(t) - A^n(\zeta^n(t))).$$

Next, let us consider the case $\zeta^n(t) > 0$. Since each customer arriving before time $\zeta^n(t)$ will either have entered service or abandoned the system by time t,

$$\hat{N}^n(t)^+ \le E^n(t) - E^n(\zeta^n(t)) + \Delta^n(\zeta^n(t))$$

where $\Delta^n(t) = E^n(t) - E^n(t-)$ is the number of customers who arrive (exactly) at time t. By (3.31), $\zeta^n(t) \le t$ and then

$$\sup_{0 \le t \le T} \Delta^n(\zeta^n(t)) \le \|\Delta^n\|_T,$$

for $\|\Delta^n\|_T = \sup_{0\le t\le T} \Delta^n(t)$. Thus,

$$\hat{N}^n(t)^+ \le E^n(t) - E^n(\zeta^n(t)) + \|\Delta^n\|_T. \tag{4.46}$$

Similarly, because a customer who arrives during $(\zeta^n(t), t]$ will either be waiting in the buffer at time t or has abandoned the system by t, one has

$$\hat{N}^n(t)^+ \ge E^n(t) - E^n(\zeta^n(t)) - (A^n(t) - A^n(\zeta^n(t))). \tag{4.47}$$

Because the customers who arrive before time $\zeta^n(t)$ cannot be waiting in the buffer at time t, for $j = 1, \ldots, d$,

$$Q_j^n(t) \le \Psi_j^0(\hat{N}^n(0)^+ + E^n(t)) - \Psi_j^0(\hat{N}^n(0)^+ + E^n(\zeta^n(t)) - \|\Delta^n\|_T). \tag{4.48}$$

Similarly, the customers who arrive during $(\zeta^n(t), t]$ cannot get into service by time t. Then,

$$Q_j^n(t) \ge \Psi_j^0(\hat{N}^n(0)^+ + E^n(t)) - \Psi_j^0(\hat{N}^n(0)^+ + E^n(\zeta^n(t))) - (A^n(t) - A^n(\zeta^n(t))). \tag{4.49}$$

Combining (4.46)–(4.49), we have

$$\check{\Lambda}_j^n(t) \le Q_j^n(t) - p_j \hat{N}^n(t)^+ \le \check{\Pi}_j^n(t), \tag{4.50}$$

where

$$\begin{aligned}
\check{\Lambda}_j^n(t) &= \hat{\Psi}_j^0(\hat{N}^n(0)^+ + E^n(t)) - \hat{\Psi}_j^0(\hat{N}^n(0)^+ + E^n(\zeta^n(t))) \\
&\quad - (A^n(t) - A^n(\zeta^n(t))) - \|\Delta^n\|_T, \\
\check{\Pi}_j^n(t) &= \hat{\Psi}_j^0(\hat{N}^n(0)^+ + E^n(t)) - \hat{\Psi}_j^0((\hat{N}^n(0))^+ + E^n(\zeta^n(t)) - \|\Delta^n\|_T) \\
&\quad + A^n(t) - A^n(\zeta^n(t)) + \|\Delta^n\|_T.
\end{aligned}$$

Combining (4.45) and (4.50), we have

$$\Lambda_j^n(t) \wedge \check{\Lambda}_j^n(t) \le Q_j^n(t) - p_j \hat{N}^n(t)^+ \le \Pi_j^n(t) \vee \check{\Pi}_j^n(t).$$

We prove this theorem by showing

$$\frac{1}{\sqrt{n}}(\Lambda_j^n, \check{\Lambda}_j^n, \Pi_j^n, \check{\Pi}_j^n) \Rightarrow 0 \quad \text{as } n \to \infty.$$

Recall that $\bar{K}^n(0) = \hat{N}^n(0)/n$. Then,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\Lambda_j^n(t) &= \tilde{\Psi}_j^{0,n}(\bar{K}^n(0)^+ + \bar{E}^n(t)) - \tilde{\Psi}_j^{0,n}(\bar{B}^n(t)) - (\tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t))), \\
\frac{1}{\sqrt{n}}\Pi_j^n(t) &= \tilde{\Psi}_j^{0,n}(\bar{K}^n(0)^+ + \bar{E}^n(t)) - \tilde{\Psi}_j^{0,n}(\bar{B}^n(t)) + (\tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t))).
\end{aligned}$$

By (4.41), Lemma 4.5, and the random-time-change theorem,

$$\sup_{0\le t\le T} \left|\tilde{\Psi}_j^{0,n}(\bar{K}^n(0)^+ + \bar{E}^n(t)) - \tilde{\Psi}_j^{0,n}(\bar{B}^n(t))\right| \Rightarrow 0 \quad \text{as } n \to \infty.$$

This, along with Lemma 4.7, implies that

$$\frac{1}{\sqrt{n}}(\Lambda_j^n, \Pi_j^n) \Rightarrow 0 \quad \text{as } n \to \infty.$$

Note that

$$\begin{aligned}
\frac{1}{\sqrt{n}}\check{\Lambda}_j^n(t) &= \tilde{\Psi}_j^{0,n}\left((\bar{K}^n(0))^+ + \bar{E}^n(t)\right) - \tilde{\Psi}_j^{0,n}\left((\bar{K}^n(0))^+ + \bar{E}^n(\zeta^n(t))\right) \\
&\qquad - (\tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t))) - \|\tilde{\Delta}^n\|_T, \\
\frac{1}{\sqrt{n}}\check{\Pi}_j^n(t) &= \tilde{\Psi}_j^{0,n}\left((\bar{K}^n(0))^+ + \bar{E}^n(t)\right) - \tilde{\Psi}_j^{0,n}\left((\bar{K}^n(0))^+ + \bar{E}^n(\zeta^n(t)) - \|\bar{\Delta}^n\|_T\right) \\
&\qquad + \tilde{A}^n(t) - \tilde{A}^n(\zeta^n(t)) + \|\tilde{\Delta}^n\|_T.
\end{aligned}$$

By Lemmas 3.8 and 4.5 and the random-time-change theorem,

$$\bar{E}^n \circ \zeta^n \Rightarrow \bar{E} \quad \text{as } n \to \infty.$$

Using (4.5),

$$\frac{1}{\sqrt{n}}\|\Delta^n\|_T \Rightarrow 0 \quad \text{as } n \to \infty.$$

Then, by similar arguments, we deduce that

$$\frac{1}{\sqrt{n}}(\check{\Lambda}_j^n, \check{\Pi}_j^n) \Rightarrow 0 \quad \text{as } n \to \infty,$$

which completes the proof. □

4.7 Proof of Theorem 4.4

Proof of Theorem 4.4. By Lemmas 2.2 and 2.3, all customers arriving prior to $t \geq 0$ will have either got into service or abandoned the system by $t + W^n(t)$. Then,

$$\hat{N}^n(t + W^n(t))^+ \leq E^n(t + W^n(t)) - E^n(t).$$

For a customer who arrives during $(t, t + W^n(t)]$, he can possibly be waiting in the buffer at time $t + W^n(t)$, or have abandoned the system by $t + W^n(t)$, or starts his service (exactly) at $t + W^n(t)$. Therefore,

$$E^n(t + W^n(t)) - E^n(t) \leq \hat{N}^n(t + W^n(t))^+ + A^n(t + W^n(t)) - A^n(t) + \Delta_D^n(t + W^n(t)),$$

where $\Delta_D^n(t) = D^n(t) - D^n(t-)$ is the number of service completions (exactly) at time t. Then by (4.6),

$$\begin{aligned} 0 &\leq \frac{1}{\sqrt{n}}\lambda^n W^n(t) - \tilde{N}^n(t + W^n(t))^+ + \tilde{E}^n(t + W^n(t)) - \tilde{E}^n(t) \\ &\leq \tilde{A}^n(t + W^n(t)) - \tilde{A}^n(t) + \tilde{\Delta}_D^n(t + W^n(t)), \end{aligned}$$

where $\tilde{\Delta}_D^n(t) = \Delta_D^n(t)/\sqrt{n}$. This leads to

$$\begin{aligned} &\left|\mu\sqrt{n}W^n(t) - \tilde{N}^n(t + W^n(t))^+\right| \leq \left|\sqrt{n}\Big(\frac{1}{n}\lambda^n - \mu\Big)W^n(t)\right| \\ &\quad + \left|\tilde{E}^n(t + W^n(t)) - \tilde{E}^n(t)\right| + \left|\tilde{A}^n(t + W^n(t)) - \tilde{A}^n(t)\right| + \tilde{\Delta}_D^n(t + W^n(t)). \quad (4.51) \end{aligned}$$

Next we show that all terms on the right-hand side of (4.51) converge in distribution to zero as $n \to \infty$. By Proposition 3.6, we have

$$W^n \Rightarrow 0 \quad \text{as } n \to \infty. \tag{4.52}$$

Using (1.1), we get

$$\left|\sqrt{n}\Big(\frac{1}{n}\lambda^n - \mu\Big)W^n\right| \Rightarrow 0 \quad \text{as } n \to \infty.$$

For any $T > 0$, by (4.5) and (4.52),

$$\sup_{0 \leq t \leq T} \left|\tilde{E}^n(t + W^n(t)) - \tilde{E}^n(t)\right| \Rightarrow 0 \quad \text{as } n \to \infty.$$

By Lemma 4.6 and (4.52),

$$\sup_{0\le t\le T}\left|\tilde{A}^n(t+W^n(t))-\tilde{A}^n(t)\right|\Rightarrow 0 \quad \text{as } n\to\infty.$$

Set $\tilde{D}^n(t)=(D^n(t)-n\mu t)/\sqrt{n}$. It follows from (4.26) that $\tilde{D}^n\Rightarrow\tilde{D}$ as $n\to\infty$, where

$$\tilde{D}(t)=-e'\tilde{M}(t)+e'R\int_0^t\tilde{Z}(s)\,\mathrm{d}s.$$

Since $\tilde{D}$ is continuous almost surely, using (4.52) again, we have

$$\sup_{0\le t\le T}\left|\tilde{\Delta}_D^n(t+W^n(t))\right|\Rightarrow 0 \quad \text{as } n\to\infty.$$

We deduce from (4.51) that

$$\sup_{0\le t\le T}\left|\mu\sqrt{n}W^n(t)-\tilde{N}^n(t+W^n(t))^+\right|\Rightarrow 0 \quad \text{as } n\to\infty. \tag{4.53}$$

By (4.15), the process $\tilde{N}$ is continuous almost surely. Then, so is the process $\tilde{N}^+$. By Lemma 2.3, $s+W^n(s)\le t+W^n(t)$ for $0\le s\le t$. Because the process $\tilde{N}^+$ is continuous almost surely, by (4.52) and the random-time-change theorem,

$$(\tilde{N}^n\circ(\zeta+W^n))^+\Rightarrow\tilde{N}^+ \quad \text{as } n\to\infty \tag{4.54}$$

where $\zeta(t)=t$ for $t\ge 0$. By (4.53), (4.54), and the convergence-together theorem, we finally see $\sqrt{n}W^n\Rightarrow\tilde{N}^+/\mu$ as $n\to\infty$. □

CHAPTER V

NUMERICAL ANALYSIS OF DIFFUSION MODELS

The limit theorems in Chapter 4 imply that in the QED regime, the dynamics of a $G/Ph/n+GI$ queue with many servers can be approximated by a multidimensional diffusion process. This chapter focuses on the numerical analysis of such diffusion processes. We propose two diffusion models for a $GI/Ph/n+GI$ queue (the first GI signifies a renewal arrival process). In each model, a multidimensional diffusion process is used to represent the scaled numbers of customers among service phases. The difference between the two diffusion models lies in how the patience time distribution is built into them. The first diffusion model uses the patience time density at zero and the second one uses the entire patience time distribution.

The diffusion models are obtained by replacing certain scaled renewal processes by Brownian motions. This replacement procedure can be justified by the FCLT. The first diffusion model is rooted in Theorem 4.3 of this thesis and the second one is motivated by the diffusion limit proved in [45]. In contrast to a diffusion limit that is for a sequence of queues, a diffusion model is built for a specific queue. The traffic intensity of this specific queue, rather than the limit traffic intensity $\rho = 1$, is incorporated so that the diffusion models can capture the queue's evolution even if the number of servers is not so large.

Next, we propose a numerical algorithm to solve the stationary distributions of the diffusion models. The computed stationary distribution is used to estimate the performance measures of the $GI/Ph/n+GI$ queue. The proposed algorithm is a variant of the one developed in [10], which computes the stationary distribution of an SRBM. As in [10], the starting point of our algorithm is the basic adjoint relationship

that characterizes the stationary distribution of a diffusion process. With an appropriate reference density, the algorithm can produce a stationary density that satisfies this relationship.

We set up a Hilbert space using the reference density. In this space, the stationary density is orthogonal to an infinite-dimensional subspace H. A finite-dimensional subspace H_k is used to approximate H and a function orthogonal to H_k can be numerically computed by solving a system of finitely many linear equations. This function is used to approximate the stationary density. There are two sources of error in computing the approximate stationary density: the *approximation error* and the *round-off error*. The approximation error arises because H_k is an approximation of H. As H_k increases to H, the approximation error decreases to zero. The round-off error occurs because the solution to the system of linear equations has error due to the finite precision of a computer. As H_k increases to H, the dimension of the linear system gets higher and the coefficient matrix of the linear system becomes closer to singular. As a consequence, the round-off error increases. The condition number of the matrix is used as a proxy for the round-off error. Balancing the approximation error and the round-off error is an important issue in our algorithm.

A properly chosen reference density is essential for the convergence of the algorithm. By convergence, we mean that the approximation error converges to zero as H_k increases to H. More importantly, a "good" reference density can make H_k converge to H quickly so that both the approximation error and the round-off error are small while the dimension of H_k is moderate. To ensure the convergence of the algorithm, the reference density should have a comparable or slower decay rate than the stationary density. Since the stationary density is unknown, we make a conjecture on the tail behavior of the limit queue length process of many-server queues with customer abandonment. We conjecture that the limit queue length process has a Gaussian tail and the tail depends on the service time distribution only through its first two

moments. This tail is used to construct a product-form reference density. With this reference density, the algorithm appears to converge quickly, producing stable and accurate results. For comparison purposes, we also test the algorithm with a certain "naively" chosen reference density in Section 5.6.1. The algorithm fails to converge with the "naive" reference density.

The basic adjoint relationship for a diffusion process is introduced in Section 5.1. In Section 5.2, we begin with recapitulating the generic algorithm of [10], and then propose a finite element implementation that follows [50]. Two diffusion models for $GI/Ph/n+GI$ queues are presented in Section 5.3. In Section 5.4, we discuss how to choose an appropriate reference density exploiting the tail behavior of a limit queue length process. In Section 5.5, it is demonstrated via numerical examples that the diffusion models serve as good approximations of many-server queues. Section 5.6 is dedicated to some implementation issues arising from the proposed algorithm.

5.1 Basic adjoint relationship

Consider the diffusion process X in (4.1). Assume that both b and σ are Lipschitz continuous so that the stochastic differential equation

$$X(t) = X(0) + \int_0^t b(X(s))\,\mathrm{d}s + \int_0^t \sigma(X(s))\,\mathrm{d}B(s)$$

has a unique strong solution, and that σ is uniformly elliptic, i.e., there exists a constant $c_2 > 0$ such that

$$y'\Sigma(x)y \geq c_2 y'y \quad \text{for all } x, y \in \mathbb{R}^d, \tag{5.1}$$

where

$$\Sigma(x) = \sigma(x)\sigma'(x). \tag{5.2}$$

We use certain diffusion processes to approximate the dynamics of a queue with many parallel servers. More specifically, two diffusion processes are identified to model such a queue and the coefficients b and σ will be mapped out explicitly in terms of

primitive data of the queue. The first diffusion model is rooted in Theorem 4.3 of this thesis, and the second one is rooted in Theorem 3.1 of [45].

A probability distribution π on $\mathbb{R}^d$ is said to be a stationary distribution of X if $X(t)$ follows distribution π for each $t > 0$ whenever $X(0)$ has distribution π. Condition (5.1) is required to ensure the uniqueness of the stationary distribution [13]. In this chapter, we assume that X has a unique stationary distribution π and π has a density g with respect to the Lebesgue measure on $\mathbb{R}^d$. For a general diffusion process, there is no explicit solution for π. In Section 5.2, we present a numerical algorithm for computing π. As in [10], the starting point of the algorithm is the basic adjoint relationship

$$\int_{\mathbb{R}^d} \mathcal{G}f(x)\,\pi(\mathrm{d}x) = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}^d), \tag{5.3}$$

where $\mathcal{G}$ is the generator of X defined by

$$\mathcal{G}f(x) = \sum_{j=1}^{d} b_j(x)\frac{\partial f(x)}{\partial x_j} + \frac{1}{2}\sum_{j=1}^{d}\sum_{\ell=1}^{d} \Sigma_{j\ell}(x)\frac{\partial^2 f(x)}{\partial x_j \partial x_\ell} \quad \text{for each } f \in C_b^2(\mathbb{R}^d) \tag{5.4}$$

and Σ is the covariance matrix given by (5.2). The following theorem is a consequence of Proposition 9.2 in [14].

Theorem 5.1. *Let π be a probability distribution on $\mathbb{R}^d$ that satisfies (5.3). Then, π is a stationary distribution of X.*

We conjecture that a stronger version of Theorem 5.1 is true.

Conjecture 5.2. *Let π be a signed measure on $\mathbb{R}^d$ that satisfies (5.3) and $\pi(\mathbb{R}^d) = 1$. Then, π is a nonnegative measure and consequently it is a stationary distribution of X.*

Our algorithm is to construct a function g on $\mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} g(x)\,\mathrm{d}x = 1 \quad \text{and} \quad \int_{\mathbb{R}^d} \mathcal{G}f(x)g(x)\,\mathrm{d}x = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}^d). \tag{5.5}$$

Assuming that Conjecture 5.2 is true, g must be the unique stationary density of X. As a special case, the nonnegativity of a signed measure π that satisfies (5.3) for a piecewise OU process was proposed as an open problem in [8].

5.2 A finite element algorithm

In this section, we propose a numerical algorithm computing the stationary density g. The basic algorithm follows the one developed in [10]. The finite element implementation closely follows [50].

5.2.1 Reference density

To compute the stationary density g, we adopt a notion called a *reference density* that was first introduced by [10]. A reference density for g is a function r defined from $\mathbb{R}^d$ to $\mathbb{R}_+$ such that

$$\int_{\mathbb{R}^d} r(x)\,\mathrm{d}x < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} q^2(x) r(x)\,\mathrm{d}x < \infty, \tag{5.6}$$

where

$$q(x) = \frac{g(x)}{r(x)} \quad \text{for each } x \in \mathbb{R}^d$$

is called the *ratio function*. Such a function r exists because $r = g$ is a reference density. The reference density controls the convergence of our algorithm. We will discuss how to choose a reference density for the diffusion models of a many-server queue in Section 5.4.

For the rest of Section 5.2, we assume that a reference density r satisfying (5.6) has been determined and remains fixed. In addition, we assume that

$$\int_{\mathbb{R}^d} b_j^2(x) r(x)\,\mathrm{d}x < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} \Sigma_{j\ell}^2(x) r(x)\,\mathrm{d}x < \infty \tag{5.7}$$

for $j, \ell = 1, \ldots, d$. Since both b and σ are Lipschitz continuous, condition (5.7) is satisfied if

$$\int_{\mathbb{R}^d} |x|^4\, r(x)\,\mathrm{d}x < \infty. \tag{5.8}$$

Let $L^2(\mathbb{R}^d, r)$ be the space of all square-integrable functions on $\mathbb{R}^d$ with respect to the measure that has density r, i.e.,

$$L^2(\mathbb{R}^d, r) = \left\{ f \in \mathcal{B}(\mathbb{R}^d) : \int_{\mathbb{R}^d} f^2(x) r(x)\,\mathrm{d}x < \infty \right\}$$

where $\mathcal{B}(\mathbb{R}^d)$ is the set of Borel-measurable functions on $\mathbb{R}^d$. Condition (5.6) implies that $q \in L^2(\mathbb{R}^d, r)$. We define an inner product on $L^2(\mathbb{R}^d, r)$ by

$$\langle f, \check{f} \rangle = \int_{\mathbb{R}^d} f(x)\check{f}(x)r(x)\,\mathrm{d}x \quad \text{for } f, \check{f} \in L^2(\mathbb{R}^d, r).$$

The induced norm is given by

$$\|f\| = \langle f, f \rangle^{1/2} \quad \text{for each } f \in L^2(\mathbb{R}^d, r). \tag{5.9}$$

One can check that $L^2(\mathbb{R}^d, r)$ is a Hilbert space and assumption (5.7) ensures that $\mathcal{G}f \in L^2(\mathbb{R}^d, r)$ for all $f \in C_b^2(\mathbb{R}^d)$. In $L^2(\mathbb{R}^d, r)$, the basic adjoint relationship in (5.5) is equivalent to

$$\langle \mathcal{G}f, q \rangle = 0 \quad \text{for all } f \in C_b^2(\mathbb{R}^d). \tag{5.10}$$

With a fixed reference density r, we need only compute the ratio function q by (5.10). Once q is obtained, we can compute the stationary density via $g(x) = q(x)r(x)$ for $x \in \mathbb{R}^d$.

Let

$$H = \text{the closure of } \{\mathcal{G}f : f \in C_b^2(\mathbb{R}^d)\} \tag{5.11}$$

where the closure is taken in the norm given by (5.9). As a subspace of $L^2(\mathbb{R}^d, r)$, H is orthogonal to q. Let $c(x) = 1$ for all $x \in \mathbb{R}^d$. Clearly, $c \in L^2(\mathbb{R}^d, r)$ but $c \notin H$ because

$$\langle c, q \rangle = \int_{\mathbb{R}^d} g(x)\,\mathrm{d}x = 1. \tag{5.12}$$

Let

$$\bar{c} = \arg\min_{f \in H} \|c - f\| \tag{5.13}$$

be the projection of c onto H. Then, $c - \bar{c}$ must be orthogonal to H. Assuming that Conjecture 5.2 holds and X has a unique stationary density g, one must have $q = \kappa_q(c - \bar{c})$ for some constant $\kappa_q \in \mathbb{R}$. By (5.12), the normalizing constant κ_q satisfies

$$\kappa_q^{-1} = \langle c, c - \bar{c} \rangle = \langle c - \bar{c}, c - \bar{c} \rangle + \langle \bar{c}, c - \bar{c} \rangle = \|c - \bar{c}\|^2.$$

Hence, the ratio function is given by

$$q = \frac{c - \bar{c}}{\|c - \bar{c}\|^2}. \tag{5.14}$$

5.2.2 An approximate stationary density

To compute q using (5.14), we need first compute $\bar{c}$, the projection of c onto H. The space H is linear and infinite-dimensional (i.e., a basis of H contains infinitely many functions). In general, solving (5.13) in an infinite-dimensional space is impossible. In the algorithm, we use a finite-dimensional subspace H_k to approximate H.

Suppose that there exists a sequence of finite-dimensional subspaces $\{H_k : k \in \mathbb{N}\}$ of H such that $H_k \to H$ in $L^2(\mathbb{R}^d, r)$ as $k \to \infty$. Here, $H_k \to H$ in $L^2(\mathbb{R}^d, r)$ means that for each $f \in H$, there exists a sequence of functions $\{\varphi_k : k \in \mathbb{N}\}$ with $\varphi_k \in H_k$ such that $\|\varphi_k - f\| \to 0$ as $k \to \infty$. Let

$$\bar{c}_k = \arg\min_{f \in H_k} \|c - f\|$$

be the projection of c onto H_k. By Proposition 7 of [10], we have the following approximation result.

Proposition 5.3. *Assume that Conjecture 5.2 is true and $H_k \to H$ in $L^2(\mathbb{R}^d, r)$. Then,*

$$\|q_k - q\| \to 0 \quad \text{as } k \to \infty,$$

where $q_k = (c - \bar{c}_k)/\|c - \bar{c}_k\|^2$. Moreover, when the reference density r is bounded,

$$\int_{\mathbb{R}^d} (g_k(x) - g(x))^2 \, \mathrm{d}x \to 0 \quad \text{as } k \to \infty,$$

where $g_k(x) = q_k(x)r(x)$ for each $x \in \mathbb{R}^d$.

As in [10], we choose

$$H_k = \{\mathcal{G}f : f \in C_k\} \tag{5.15}$$

for some finite-dimensional space C_k. In Section 5.2.3, we will discuss how to construct C_k using a finite element method. For notational convenience, we omit the subscript k when k is fixed. The finite-dimensional functional space is thus denoted by C. Let m_C be the dimension of C and $\{f_i : i = 1, \ldots, m_C\}$ be a basis of C. We assume that the family $\{\mathcal{G}f_i : i = 1, \ldots, m_C\}$ is linearly independent in $L^2(\mathbb{R}^d, r)$. Then,

$$\bar{c}_k = \sum_{i=1}^{m_C} u_i \mathcal{G} f_i \quad \text{for some } u_i \in \mathbb{R} \text{ and } i = 1, \ldots, m_C. \tag{5.16}$$

Using the fact $\langle \mathcal{G}f_i, c - \bar{c}_k \rangle = 0$ for $i = 1, \ldots, m_C$, we obtain a system of linear equations

$$Au = v \tag{5.17}$$

where

$$A_{i\ell} = \langle \mathcal{G}f_i, \mathcal{G}f_\ell \rangle, \quad u = (u_1, \ldots, u_{m_C})', \quad v_i = \langle \mathcal{G}f_i, c \rangle. \tag{5.18}$$

By the linear independence assumption, the $m_C \times m_C$ matrix A is positive definite. Thus, $u = A^{-1}v$ is the unique solution to (5.17). Once the vector u is obtained, we can compute the projection $\bar{c}_k$ by (5.16). Finally, the stationary density g can be approximated via

$$g(x) \approx g_k(x) = r(x) \frac{c(x) - \bar{c}_k(x)}{\|c - \bar{c}_k\|^2} \quad \text{for each } x \in \mathbb{R}^d.$$

5.2.3 A finite element method

In [10], the authors employed polynomials of orders up to k to construct the space C_k. This choice appears to be numerically unstable. The approximation error is significant when k is small, say, $k \leq 5$. As k increases, the round-off error in solving (5.17) increases and ultimately dominates the approximation error. Although their implementation produces accurate estimates for the stationary means of SRBMs, it sometimes produces poor estimates for the stationary distributions. In this section, we construct a sequence of spaces $\{C_k : k \in \mathbb{N}\}$ using the finite element method as

in [50]. Because the state space in [50] is bounded, neither a reference density nor state space truncation is used there.

The state space of X is unbounded in our setting. It is necessary to truncate the state space to apply the finite element method. Let $\{K_k : k \in \mathbb{N}\}$ be a sequence of compact sets in $\mathbb{R}^d$. For each $f \in C_k$, we assume that $f(x) = 0$ for $x \in \mathbb{R}^d \setminus K_k$. The subscript k is omitted again when it is fixed and we use K to denote the compact support of the space C. In our implementation, we restrict K to be a d-dimensional hypercube

$$K = [-\zeta_1, \xi_1] \times \cdots \times [-\zeta_d, \xi_d], \tag{5.19}$$

where both ζ_j and ξ_j are positive constants for $j = 1, \ldots, d$.

We partition K into a finite number of subdomains. Such a partition is called a *mesh* and each subdomain is called a *finite element*. Since K is a hypercube, it is natural to use a lattice mesh, where each finite element is again a hypercube. In this case, each corner point of a finite element is called a *node*. In dimension $j = 1, \ldots, d$, we divide the interval $[-\zeta_j, \xi_j]$ into n_j subintervals by partition points

$$-\zeta_j = y_j^0 < y_j^1 < \cdots < y_j^{n_j} = \xi_j.$$

Then, K is divided into $\prod_{j=1}^d n_j$ finite elements. For future reference, we label the nodes following the way that node $(i_1, \ldots, i_d)$ corresponds to spatial coordinate $(y_1^{i_1}, \ldots, y_d^{i_d})$, and define

$$h_j^\ell = y_j^{\ell+1} - y_j^\ell \quad \text{for } \ell = 0, \ldots, n_j - 1 \text{ and } j = 1, \ldots, d.$$

If Δ denotes such a mesh, we define

$$|\Delta| = \max\{h_j^\ell : \ell = 0, \ldots, n_j - 1;\ j = 1, \ldots, d\}$$

and

$$\eta_\Delta = \max\left\{\frac{h_{j_1}^{\ell_1}}{h_{j_2}^{\ell_2}} : \ell_1, \ell_2 = 0, \ldots, n_j - 1;\ j_1, j_2 = 1, \ldots, d;\ j_1 \neq j_2\right\}. \tag{5.20}$$

The finite-dimensional space C is generated using the above mesh. We use the cubic Hermite basis functions to construct a basis of C, as in [50]. The one-dimensional Hermite basis functions for $-1 \le z \le 1$ are given by

$$\phi(z) = (|z| - 1)^2(2\,|z| + 1) \quad \text{and} \quad \psi(z) = z(|z| - 1)^2. \tag{5.21}$$

In dimension $j = 1, \ldots, d$ and for $\ell = 1, \ldots, n_j - 1$, let

$$\phi_j^\ell(z) = \begin{cases} \phi\left(\dfrac{z - y_j^\ell}{h_j^{\ell-1}}\right) & \text{if } y_j^{\ell-1} \le z \le y_j^\ell, \\ \phi\left(\dfrac{z - y_j^\ell}{h_j^\ell}\right) & \text{if } y_j^\ell \le z \le y_j^{\ell+1}, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\psi_j^\ell(z) = \begin{cases} h_j^{\ell-1}\psi\left(\dfrac{z - y_j^\ell}{h_j^{\ell-1}}\right) & \text{if } y_j^{\ell-1} \le z \le y_j^\ell, \\ h_j^\ell\psi\left(\dfrac{z - y_j^\ell}{h_j^\ell}\right) & \text{if } y_j^\ell \le z \le y_j^{\ell+1}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $x = (x_1, \ldots, x_d)'$ be a vector in K. At node $(i_1, \ldots, i_d)$, the basis functions of C are the tensor-product Hermite basis functions

$$f_{i_1,\ldots,i_d,\chi_1,\ldots,\chi_d}(x) = \prod_{j=1}^{d} g_{i_j,\chi_j}(x_j) \tag{5.22}$$

where χ_j is either 0 or 1 and

$$g_{i_j,\chi_j}(z) = \begin{cases} \phi_j^{i_j}(z) & \text{if } \chi_j = 0, \\ \psi_j^{i_j}(z) & \text{if } \chi_j = 1. \end{cases}$$

Therefore, each node has 2^d tensor-product basis functions and the space C has a total of

$$m_C = 2^d \prod_{j=1}^{d} (n_j - 1) \tag{5.23}$$

basis functions.

The space C is not a subspace of $C_b^2(\mathbb{R}^d)$. For the one-dimensional Hermite basis functions in (5.21), the second order derivative of $\phi(z)$ is not defined at $z = -1, 1$ and the second order derivative of $\psi(z)$ is not defined at $z = -1, 0, 1$. As a consequence, there exists $f \in C$ for which $\mathcal{G}f$ is not defined on the boundaries of certain finite elements. Because such boundaries have Lebesgue measure zero in $\mathbb{R}^d$, for each $f \in C$, we can find a sequence of functions $\{\varphi_i : i \in \mathbb{N}\}$ in $C_b^2(\mathbb{R}^d)$ such that $\|\mathcal{G}\varphi_i - \mathcal{G}f\| \to 0$ as $i \to \infty$. Hence, $H_k \subset H$ still holds for each k.

For the linear system (5.17) to have a unique solution, the family of functions

$$\{\mathcal{G}f_{i_1,\ldots,i_d,\chi_1,\ldots,\chi_d} : i_j = 1, \ldots, n_j - 1; \chi_j = 0, 1; j = 1, \ldots d\}$$

must be linearly independent in $L^2(\mathbb{R}^d, r)$. The following proposition provides sufficient conditions for the linear independence. Its proof can be found in Appendix B.

Proposition 5.4. *Let $\mathcal{G}$ be the generator of X in (5.4) such that conditions (4.2) and (5.1) hold and all entries of Σ are continuously differentiable. Assume that $r(x) > 0$ for all $x \in \mathbb{R}^d$. Then, the family of functions*

$$\{\mathcal{G}f_{i_1,\ldots,i_d,\chi_1,\ldots,\chi_d} : i_j = 1, \ldots, n_j - 1; \chi_j = 0, 1; j = 1, \ldots d\}$$

are linearly independent in $L^2(\mathbb{R}^d, r)$, where $f_{i_1,\ldots,i_d,\chi_1,\ldots,\chi_d}$ is the basis function of C given by (5.22). Consequently, the solution to the linear system (5.17) is unique.

Now let us consider a sequence of functional spaces $\{C_k : k \in \mathbb{N}\}$. Let Δ_k be the mesh for constructing C_k. We assume that the mesh Δ_{k+1} is a refinement of Δ_k, i.e., a node or an interelement boundary in Δ_k is also a node or an interelement boundary in Δ_{k+1}. We further assume that such refinements are *regular*, i.e., for each η_{Δ_k} defined in (5.20), the set $\{\eta_{\Delta_k} : k \in \mathbb{N}\}$ is bounded. The next proposition, along with Proposition 5.3, justifies the proposed algorithm for computing the stationary distribution. We leave the proof of Proposition 5.5 to Appendix C.

Proposition 5.5. *Let $\{\Delta_k : k \in \mathbb{N}\}$ be a sequence of lattice meshes such that each Δ_{k+1} is a refinement of Δ_k and the refinements are regular. Let K_k be the d-dimensional finite hypercube that is the domain of Δ_k, and C_k be the functional space generated by Δ_k using the tensor-product Hermite basis functions in (5.22). Let H be the infinite-dimensional space in (5.11) and H_k be the finite-dimensional space in (5.15), where the generator $\mathcal{G}$ satisfies (4.2) and (5.7). Assume that*

$$|\Delta_k| \to 0 \quad \textit{and} \quad K_k \uparrow \mathbb{R}^d \quad \textit{as } k \to \infty.$$

Then,

$$H_k \to H \quad \textit{as } k \to \infty.$$

5.3 Diffusion models for $GI/Ph/n + GI$ queues

In this section, two diffusion models are presented for a $GI/Ph/n + GI$ queue in the QED regime. We focus on a queue with a fixed number of servers, as opposed to a sequence of queues indexed by their server numbers. Hence, n is fixed and the QED regime should be interpreted in the following sense: In this queue, both the arrival rate λ and the number of servers n are large, and the traffic intensity $\rho = \lambda/(n\mu)$ is close to one. Because customer abandonment is allowed, it is not necessary to assume $\rho < 1$ for the system to reach a steady state. For future purposes, we put

$$\beta_0 = \sqrt{n}(1 - \rho). \tag{5.24}$$

Recall that at time t and for $j = 1, \ldots, d$, $Z_j(t)$ is the number of customers in phase j service, $\mathcal{Q}_j(t)$ is the number of waiting customers whose service begins with phase j, and

$$Y_j(t) = Z_j(t) + \mathcal{Q}_j(t) \tag{5.25}$$

is the number of phase j customers in system. Here, we no longer attach the superscript n to index the processes since n is fixed. Let $Y(t)$ be the corresponding

d-dimensional random vector and

$$\breve{Y}(t) = \frac{1}{\sqrt{n}}(Y(t) - n\theta) \tag{5.26}$$

where θ is the work load distribution defined by (4.9). In each diffusion model, the process $\breve{Y}$ is approximated by a d-dimensional diffusion process.

For $j = 1, \ldots, d$, let $\mathcal{A}_j(t)$ be the cumulative number of phase j customers who have abandoned the system by time t and $\mathcal{A}(t)$ be the corresponding d-dimensional random vector. One can check that the process Y satisfies the following equation

$$Y(t) = Y(0) + \Phi^0(E(t)) + \sum_{j=1}^{d} \Phi^j(S_j(T_j(t))) - S(T(t)) - \mathcal{A}(t), \tag{5.27}$$

where E is the renewal arrival process of the queue, $S_1, \ldots, S_d, \Phi^0, \ldots, \Phi^d$ are the processes defined in Section 4.4, and $S(T(t)) = (S_1(T_1(t)), \ldots, S_d(T_d(t)))'$ with

$$T_j(t) = \int_0^t Z_j(s)\,\mathrm{d}s. \tag{5.28}$$

To derive the diffusion models, consider a scaled version of (5.27). We define several scaled processes by

$$\breve{E}(t) = \frac{1}{\sqrt{n}}(E(t) - \lambda t), \quad \breve{S}(t) = \frac{1}{\sqrt{n}}(S(nt) - n\nu t), \quad \breve{Z}(t) = \frac{1}{\sqrt{n}}(Z(t) - n\theta),$$
$$\breve{\mathcal{A}}(t) = \frac{1}{\sqrt{n}}\mathcal{A}(t), \quad \breve{\Phi}^0(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor}(\phi^0(i) - p), \quad \breve{\Phi}^j(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt \rfloor}(\phi^j(i) - p^j)$$

for $j = 1, \ldots, d$, where p^j is the jth column of P'. By (5.24)–(5.26) and (5.28), the dynamical equation in (5.27) turns out to be

$$\begin{aligned}\breve{Y}(t) = \breve{Y}(0) - \beta_0\mu p t + p\breve{E}(t) + \breve{\Phi}^0\Big(\frac{E(t)}{n}\Big) \\ + \sum_{j=1}^{d}\breve{\Phi}^j\Big(\frac{S_j(T_j(t))}{n}\Big) - (I - P')\breve{S}\Big(\frac{T(t)}{n}\Big) - R\int_0^t \breve{Z}(s)\,\mathrm{d}s - \breve{\mathcal{A}}(t).\end{aligned} \tag{5.29}$$

In both diffusion models, we replace the scaled primitive processes in (5.29) by certain Brownian motions. These approximations can be justified by the FCLT. Let

B_E be a one-dimensional driftless Brownian motion with variance $\lambda c_a^2/n$, where c_a^2 is the squared coefficient of variation for the interarrival time distribution. Let $B_0, \ldots, B_d$, and B_S are d-dimensional driftless Brownian motions with covariance matrices $H^0, \ldots, H^d$, and diag(ν), respectively, with $H^0, \ldots, H^d$ given in (4.12). We assume that $\breve{Y}(0), B_E, B_0, \ldots, B_d, B_S$ are mutually independent. In the diffusion models, the above Brownian motions take the places of the scaled primitive processes $\breve{E}, \breve{\Phi}^0, \ldots, \breve{\Phi}^d, \breve{S}$, respectively. Recall that $Q(t)$ is the queue length at time t. Let

$$\breve{Q}(t) = \frac{1}{\sqrt{n}} Q(t).$$

Then, $Q(t) = (e'Y(t) - n)^+$ or equivalently,

$$\breve{Q}(t) = (e'\breve{Y}(t))^+. \tag{5.30}$$

By Theorem 4.2, these waiting customers are approximately distributed among the d phases according to distribution p when n is large, i.e.,

$$\mathcal{Q}_j(t) \approx p_j Q(t) \quad \text{for } j = 1, \ldots, d.$$

It follows from (5.25) that

$$Z(t) \approx Y(t) - pQ(t).$$

By (5.26) and (5.30), this approximation has a scaled version

$$\breve{Z}(t) \approx \breve{Y}(t) - p(e'\breve{Y}(t))^+. \tag{5.31}$$

Recall that $A(t)$ is the cumulative number of customers who have abandoned the system by time t and

$$\breve{A}(t) = \frac{1}{\sqrt{n}} A(t).$$

One can expect that these abandoned customers are also approximately distributed among the d phases by distribution p, i.e.,

$$\breve{\mathcal{A}}(t) \approx p\breve{A}(t). \tag{5.32}$$

We also exploit the following approximations

$$\frac{E(t)}{n} \approx \frac{\lambda t}{n} = \rho\mu t, \quad \frac{T(t)}{n} \approx (\rho \wedge 1)\theta t, \quad \frac{S_j(T_j(t))}{n} \approx (\rho \wedge 1)\nu_j\theta_j t. \tag{5.33}$$

The approximations in (5.31)–(5.33) are used in both diffusion models. These two models differ only in how to approximate the scaled abandonment process $\breve{A}$.

5.3.1 Diffusion model using the patience time density at zero

In the first diffusion model, the patience time distribution is used only through its density at zero when approximating $\breve{A}$. We assume that conditions (3.4) and (3.5) hold. In this model, $\breve{A}$ is approximated by

$$\breve{A}(t) \approx \alpha \int_0^t (e'\breve{Y}(s))^+ \, \mathrm{d}s \quad \text{for } t \geq 0. \tag{5.34}$$

This approximation can by justified by Theorem 3.1. When $\alpha = 0$, this approximation yields $\breve{A}(t) \approx 0$ for all $t \geq 0$. In this case, the diffusion model is for a $GI/Ph/n$ queue without abandonment.

In the dynamical equation (5.29), when the scaled primitive processes are replaced by appropriate Brownian motions and the approximations in (5.31)–(5.34) are employed, we obtain the following stochastic differential equation

$$\begin{aligned} X(t) = X(0) - \beta_0\mu p t + pB_E(t) + B_0(\rho\mu t) + \sum_{j=1}^{d} B_j((\rho \wedge 1)\nu_j\theta_j t) \\ - (I - P')B_S((\rho \wedge 1)\theta t) - R\int_0^t (X(s) - p(e'X(s))^+) \, \mathrm{d}s - p\alpha \int_0^t (e'X(s))^+ \, \mathrm{d}s \end{aligned} \tag{5.35}$$

where we take $X(0) = \breve{Y}(0)$. We may write (5.35) into the standard form (4.1) where for $x \in \mathbb{R}^d$, the drift coefficient b is

$$b(x) = -\beta_0\mu p - R(x - p(e'x)^+) - p\alpha(e'x)^+, \tag{5.36}$$

the diffusion coefficient σ is a $d \times d$ constant matrix satisfying

$$\begin{aligned}\Sigma(x) &= \sigma(x)\sigma'(x) \\ &= \rho\mu(c_a^2 pp' + H^0) \\ &\quad + (\rho \wedge 1)\bigg(\sum_{j=1}^{d} \nu_j \theta_j H^j + (I - P') \operatorname{diag}(\nu) \operatorname{diag}(\theta)(I - P)\bigg),\end{aligned} \tag{5.37}$$

and B is a d-dimensional standard Brownian motion. One can check that $\Sigma(x)$ is positive definite and thus uniformly elliptic. Both b and σ are Lipschitz continuous, so a strong solution to (5.35) exists. More specifically, the drift coefficient b in (5.36) is a piecewise linear function of x and the diffusion process X is a d-dimensional piecewise OU process. In this model, the diffusion process X depends on the patience time distribution only through its density at zero. When using the proposed algorithm to solve the stationary density, it follows from Proposition 5.4 that the linear system (5.17) has a unique solution.

If we replace ρ by one in (5.35), the resulting diffusion process turns out to be the diffusion limit for $G/Ph/n + GI$ queues in Theorem 4.3. Since ρ is close to one in the current setting, this theorem justifies the diffusion model in (5.35).

5.3.2 Diffusion model using patience time hazard rate scaling

When the patience time distribution does not have a density at zero, the diffusion model in (5.35) fails to exist. When $\alpha = 0$ and $\rho > 1$, the diffusion process X in (5.35) does not have a stationary distribution. In this case, the model cannot be a satisfactory approximation of the many-server queue, as the queue may have a stationary distribution thanks to customer abandonment. It is also demonstrated in [45] that when the density near zero rapidly changes, the system performance can be sensitive to the patience time distribution in a neighborhood of the origin. In this case, using the patience time density at zero solely may not yield adequate approximation to the queue. Our second diffusion model exploits the idea of scaling

the patience time hazard rate function, which was first proposed in [46] for single-server queues and was recently extended to many-server queues in [45].

In this model, we assume that the patience time distribution F satisfies

$$F(0) = 0$$

and it has a bounded hazard rate function h, given by

$$h(t) = \frac{f_F(t)}{1 - F(t)} \quad \text{for } t \geq 0,$$

where f_F is the density of the patience time distribution F. With the hazard rate function, F can be written by

$$F(t) = 1 - \exp\Big(- \int_0^t h(s)\,\mathrm{d}s \Big) \quad \text{for } t \geq 0.$$

In the second diffusion model, the scaled abandonment process $\breve{A}$ is approximated by

$$\breve{A}(t) \approx \int_0^t \int_0^{(e'\breve{Y}(s))^+} h\Big(\frac{\sqrt{n}u}{\lambda}\Big)\,\mathrm{d}u\,\mathrm{d}s \quad \text{for } t \geq 0. \tag{5.38}$$

The entire patience time distribution is built into the approximation through its hazard rate function. The intuition of the hazard rate scaling approximation was explained in [46]: Consider the $Q(s)$ waiting customers in the buffer at time s. In general, only a small fraction of customers can abandon the system when the queue is working in the QED regime. Then by time s, the ith customer from the back of the queue has been waiting around i/λ time units. Approximately, this customer will abandon the queue during the next δ time units with probability $h(i/\lambda)\delta$. It follows that for the system, the instantaneous abandonment rate at time s is close to $\sum_{i=1}^{Q(s)} h(i/\lambda)$. By (5.26) and (5.30), the scaled abandonment rate can be approximated by

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{Q(s)} h\Big(\frac{i}{\lambda}\Big) \approx \int_0^{\breve{Q}(s)} h\Big(\frac{\sqrt{n}u}{\lambda}\Big)\,\mathrm{d}u = \int_0^{(e'\breve{Y}(s))^+} h\Big(\frac{\sqrt{n}u}{\lambda}\Big)\,\mathrm{d}u, \tag{5.39}$$

from which (5.38) follows. Note that the arrival rate λ is in the order of $O(n)$ and $Q(s)$ is in the order of $O(n^{1/2})$. The patience time distribution in a small neighborhood of

zero, not just its density at zero, is considered in the instantaneous abandonment rate in (5.39). Hence, the hazard rate scaling approximation in (5.38) is more accurate than that in (5.34). This approximation can be justified for $GI/M/n+GI$ queues by Propositions 9.1 and 9.2 in [45]. With minor modifications to the proofs, these two propositions can be extended to $GI/Ph/n+GI$ queues.

Let m be a nonnegative integer. Suppose that the hazard rate function h is m times continuously differentiable in a neighborhood of zero. By Taylor's theorem,

$$h(z) \approx h(0) + \sum_{\ell=1}^{m} h^{(\ell)}(0)\frac{z^\ell}{\ell!}$$

for $z > 0$ small enough, where $h^{(\ell)}$ is the ℓth order derivative of h. In this case, the approximation in (5.38) turns out to be

$$\tilde{\mathcal{A}}(t) \approx h(0)\int_0^t (e'\breve{Y}(s))^+ \,\mathrm{d}s + \sum_{\ell=1}^{m} \frac{n^{\ell/2}h^{(\ell)}(0)}{\lambda^\ell(\ell+1)!}\int_0^t ((e'\breve{Y}(s))^+)^{\ell+1}\,\mathrm{d}s.$$

Because $h(0)$ is identical to the patience time density at zero, the approximation in (5.34) can be regarded as the zeroth degree Taylor's approximation of (5.38). When the patience times are exponentially distributed, the hazard rate function is constant and the two approximations in (5.34) and (5.38) are identical.

Using the Brownian motion replacement and the approximations in (5.31)–(5.33) and (5.38), we obtain the second diffusion model for the $GI/Ph/n+GI$ queue,

$$\begin{aligned} X(t) = X(0) &- \beta_0\mu p t + pB_E(t) + B_0(\rho\mu t) + \sum_{j=1}^{d} B_j((\rho\wedge 1)\nu_j\theta_j t) \\ &- (I-P')B_S((\rho\wedge 1)\theta t) - R\int_0^t (X(s) - p(e'X(s))^+)\,\mathrm{d}s \\ &- p\int_0^t \int_0^{(e'X(s))^+} h\Big(\frac{\sqrt{n}u}{\lambda}\Big)\,\mathrm{d}u\,\mathrm{d}s. \end{aligned} \tag{5.40}$$

The diffusion process X in (5.40) has the same diffusion coefficient σ as in the first model (5.35). Its drift coefficient b is

$$b(x) = -\beta_0\mu p - R(x - p(e'x)^+) - p\int_0^{(e'x)^+} h\Big(\frac{\sqrt{n}u}{\lambda}\Big)\,\mathrm{d}u \quad \text{for } x\in\mathbb{R}^d. \tag{5.41}$$

Since h is bounded, the drift coefficient b is Lipschitz continuous and the stochastic differential equation (5.40) has a strong solution. By Proposition 5.4, the solution to the linear system (5.17) is unique when we use the proposed algorithm to solve the stationary density of this diffusion model. Comparing (5.35) and (5.40), one can see that the two models differ only in how the patience time distribution is incorporated. Because a more accurate approximation is used for the abandonment process, the second model can provide a better approximation for the queue.

5.4 Choosing a reference density

The reference density controls the convergence of the proposed algorithm. In this section, we discuss how to choose appropriate reference densities for the diffusion models. Some considerations are as follows.

First, to be a reference density, a candidate function r must satisfy (5.6) even though the stationary density g is unknown. The second condition in (5.6) requires that r have a comparable or slower decay rate than g. When g is bounded, its decay rate is sufficient to determine a function r that satisfies (5.6).

Second, the most computationally expensive part of the algorithm is constructing and solving the linear system (5.17). As demonstrated by Proposition 5.5, the finite-dimensional space H_k approximates the infinite-dimensional space H better as k increases, thus reducing the approximation error. On the other hand, as the dimension of H_k increases, constructing and solving (5.17) requires more computation time and memory space. The condition number of the matrix A in (5.17) also gets worse as the dimension of H_k becomes large. This yields higher round-off error. A "good" reference density should balance the approximation error and the round-off error. With such a reference density, it is possible to have small approximation error even if the dimension of H_k is moderate.

Intuitively, when r is "close" to the stationary density g, both the ratio function

q and the projection $\bar{c}$ are close to constant. We can thus expect that a space H_k with a moderate dimension is able to produce a satisfactory approximation. All these observations motivate us to explore the tail behavior of a diffusion model.

5.4.1 Tail behavior

Let us focus on the diffusion limit in Theorem 4.3. Assume that the piecewise OU process $\tilde{X}$ is positive recurrent and has a stationary distribution. Let $\tilde{X}(\infty)$ be the corresponding d-dimensional random vector in steady state. Since ρ is close to one, the tail behavior of the diffusion process X in (5.35) is expected to be comparable to that of the limit diffusion process $\tilde{X}$.

To explore the tail behavior of $\tilde{X}(\infty)$, consider a sequence of $GI/GI/n{+}GI$ queues in the QED regime, i.e., condition (1.1) holds. If all patience times are infinite, the queues turn out to be $GI/GI/n$ queues without customer abandonment. In each queue, the service times are iid following a general distribution. We assume that these queues, each indexed by the number of servers n, have the same service time distribution.

Assume that all these queues are in steady state. Let $N^n(\infty)$ be the stationary number of customers in the nth system and

$$\tilde{N}^n(\infty) = \frac{1}{\sqrt{n}}(N^n(\infty) - n).$$

For $GI/GI/n$ queues in the QED regime, the limit queue length in steady state was studied in [16], where the service time distribution is assumed to be lattice-valued on a finite support. The authors first showed that $\tilde{N}^n(\infty)$ weakly converges to a random variable $\tilde{N}(\infty)$ as $n \to \infty$, and then proved that

$$\lim_{z\to\infty} \frac{1}{z} \log \mathbb{P}[\tilde{N}(\infty) > z] = -\frac{2\beta}{c_a^2 + c_s^2}, \tag{5.42}$$

where c_a^2 and c_s^2 are the squared coefficients of variation of the interarrival and the service time distributions, respectively. In (5.42), the decay rate does not depend on

the service time distribution beyond its first two moments. Recently, this result has been extended in [15] to $GI/GI/n$ queues with a general service time distribution.

When $\alpha = 0$ and $d = 1$, the limit diffusion process $\tilde{X}$ in Theorem 4.3 is for a sequence of $GI/M/n$ queues without customer abandonment. In this case, the service time distribution is exponential and $\tilde{N}(\infty) = \tilde{X}(\infty)$. It was proved in [20] that the stationary density of $\tilde{X}(\infty)$ has a closed-form expression

$$\tilde{g}(z) = \begin{cases} a_1 \exp\left(-\dfrac{(z+\beta)^2}{1+c_a^2} \right) & \text{if } z < 0, \\ a_2 \exp\left(-\dfrac{2\beta z}{1+c_a^2} \right) & \text{if } z \geq 0, \end{cases} \tag{5.43}$$

where a_1 and a_2 are normalizing constants making $\tilde{g}$ continuous at zero. The decay rate of $\tilde{g}$ in (5.43) is consistent with (5.42). Both formulas suggest that $\tilde{N}(\infty)$ has an exponential tail on the right side.

For a $GI/GI/n + GI$ queue with many servers and customer abandonment, the limiting tail behavior of $\tilde{N}^n(\infty)$ as $n \to \infty$ remains unknown except for very simple cases. When $\alpha > 0$ and $d = 1$, the limit diffusion process $\tilde{X}$ in Theorem 4.3 is a one-dimensional piecewise OU process. It admits a piecewise Gaussian stationary density

$$\tilde{g}(z) = \begin{cases} a_3 \exp\left(-\dfrac{(z+\beta)^2}{1+c_a^2} \right) & \text{if } z < 0, \\ a_4 \exp\left(-\dfrac{\alpha(z+\alpha^{-1}\mu\beta)^2}{\mu(1+c_a^2)} \right) & \text{if } z \geq 0, \end{cases} \tag{5.44}$$

where a_3 and a_4 are normalizing constants that make $\check{g}$ continuous at zero. See [7] for more details.

Observing (5.42) and (5.44), we conjecture that for a sequence of $GI/GI/n + GI$ queues in the QED regime, the limiting tail behavior of $\tilde{N}^n(\infty)$ depends on the service time distribution only through its first two moments, and on the patience time distribution only through its density at zero.

Conjecture 5.6. *Consider a sequence of $GI/GI/n + GI$ queues that satisfies (1.1), (3.4), and (3.5). Assume that the patience time distribution has a positive density*

at zero, i.e., $\alpha > 0$ in (3.5). Assume further that the interarrival and the service time distributions satisfy the T_0 assumptions (i)–(iii) in Section 2.1 of [15]. Then, (a) $N^n(\infty)$ exists for each n; (b) the sequence of random variables $\{\tilde{N}^n(\infty) : n \in \mathbb{N}\}$ converges in distribution to a random variable $\tilde{N}(\infty)$; (c) $\tilde{N}(\infty)$ satisfies

$$\lim_{z\to\infty} \frac{1}{z^2} \log \mathbb{P}[\tilde{N}(\infty) > z] = -\frac{\alpha}{\mu(c_a^2 + c_s^2)}. \tag{5.45}$$

The intuition below may help understand why the conjectured decay rate must be Gaussian. When $\tilde{N}(\infty) > z$ for some $z > 0$, there are more than $n^{1/2}z$ waiting customers in the queue correspondingly, and each waiting customer is "racing" to abandon the system. At any time, the instantaneous abandonment rate is approximately proportional to the queue length. In such a system, the customer departure process, including both service completions and customer abandonments, behaves as if the system is a queue with infinite servers. Thus, one can expect that the tail of the limit queue length is Gaussian, which decays much faster than an exponential tail for queues without abandonment.

5.4.2 Reference densities for model (5.35)

For $GI/Ph/n + GI$ queues, the limit diffusion process $\tilde{X}$ in Theorem 4.3 satisfies

$$\tilde{N}(\infty) = e'\tilde{X}(\infty).$$

The discussion in Section 5.4.1 gives ample evidence of the tail behavior $\mathbb{P}[\tilde{N}(\infty) > z]$ as $z \to \infty$. Although the left tail $\mathbb{P}[\tilde{N}(\infty) < -z]$ as $z \to \infty$ remains unknown when $d > 1$, our numerical experiments suggest that this tail is not sensitive to the service time distribution beyond its mean. Thus, we use the left tail for a queue with an exponential service time distribution to construct the reference density. We propose to use a product reference density

$$r(x) = \prod_{j=1}^{d} r_j(x_j) \quad \text{for } x \in \mathbb{R}^d. \tag{5.46}$$

When $\alpha = 0$ and $\rho < 1$ in (5.35), there is no abandonment in the queue. Based on (5.42) and (5.43), we choose

$$r_j(z) = \begin{cases} \exp\left(-\dfrac{(z+\theta_j\beta_0)^2}{1+c_a^2} \right) & \text{if } z < 0, \\ \exp\left(-\dfrac{2\beta_0 z}{c_a^2+c_s^2} - \dfrac{\theta_j^2\beta_0^2}{1+c_a^2} \right) & \text{if } z \geq 0, \end{cases} \tag{5.47}$$

where β_0 is given by (5.24). The function r_j has an exponential tail on the right and a Gaussian tail on the left. One can check that the reference density given by (5.46) and (5.47) satisfies condition (5.8). In (5.47), we set the shift term for $z < 0$ to be $\theta_j\beta_0$ according to the following observation. In the associated queue, β_0 is the scaled mean number of idle servers and θ_j is the fraction of phase j service load. In steady state, one can expect that $\tilde{Y}_j(t)$, the centered and scaled number of phase j customers, is around $-\theta_j\beta_0$.

When $\alpha > 0$ in (5.35), the associated queue has abandonment. By (5.44) and Conjecture 5.6, we choose

$$r_j(z) = \begin{cases} \exp\left(-\dfrac{(z+\theta_j\beta_0)^2}{1+c_a^2} \right) & \text{if } z < 0, \\ \exp\left(-\dfrac{\alpha(z+p_j\alpha^{-1}\mu\beta_0)^2}{\mu(c_a^2+c_s^2)} + \dfrac{p_j^2\alpha^{-1}\mu\beta_0^2}{c_a^2+c_s^2} - \dfrac{\theta_j^2\beta_0^2}{1+c_a^2} \right) & \text{if } z \geq 0, \end{cases} \tag{5.48}$$

whose two tails are both Gaussian but have different decay rates. This reference density also satisfies (5.8). In (5.48), the shift term for $z \geq 0$ is taken to be $p_j\mu\beta_0/\alpha$ because of the observation below. When $\rho \geq 1$, the throughput of the queue is nearly $n\mu$. Let q_0 be the scaled queue length "in equilibrium", i.e., the arrival and the departure rates of the system are balanced when the queue length is around $n^{1/2}q_0$. Because in this case the abandonment rate is $\alpha n^{1/2}q_0$, we have $\lambda = n\mu + \alpha n^{1/2}q_0$, or $q_0 = -\mu\beta_0/\alpha$ by (5.24). Since the fraction of phase j waiting customers is around p_j, $\tilde{Y}_j(t)$ is around $-p_j\mu\beta_0/\alpha$ as the queue reaches a steady state.

5.4.3 Reference densities for model (5.40)

For the diffusion model (5.40) that adopts the patience time hazard rate scaling, the tail behavior of X in steady state is left to future research. In some cases, we may exploit the diffusion limit in Theorem 4.3 to facilitate the choice of a reference density for the current model. The principle is again to ensure that the reference density has a comparable or slower decay rate than the stationary density of X. For that, we build an auxiliary queue that shares the same arrival process and service times with the $GI/Ph/n+GI$ queue, but the auxiliary queue may have no abandonment or have an exponential patience time distribution. Let $\hat{X}$ be the diffusion process in (5.35) for the auxiliary queue. If $\hat{X}$ has a slower decay rate than X, a reference density of $\hat{X}$ must be a reference density of X, too.

When $\rho < 1$, the auxiliary queue is a $GI/Ph/n$ queue. It is intuitive that the queue length decays faster in the $GI/Ph/n + GI$ queue than in the auxiliary queue because the latter has no abandonment. As a consequence, $\hat{X}$ has a slower decay rate than X and the reference density given by (5.46) and (5.47) for $\hat{X}$ can be used for the current model.

When $\rho > 1$, the auxiliary queue is a $GI/Ph/n+M$ queue. Let $\alpha > 0$ be the rate of the exponential patience time distribution, which is to be determined in order for $\hat{X}$ to have an appropriate decay rate. For that, we need investigate the abandonment process of the $GI/Ph/n + GI$ queue.

Assume that the hazard rate function h is m times continuously differentiable in a neighborhood of zero for some nonnegative integer m, and among $\ell = 0, \ldots, m$, there is at least one $h^{(\ell)}(0) \neq 0$. We follow the convention that $h^{(0)}(0) = h(0)$. Let ℓ_0 be the smallest nonnegative integer such that $h^{(\ell_0)}(0) \neq 0$. For $z > 0$ in a small neighborhood of zero, the ℓ_0th degree Taylor's approximation of h is

$$h(z) \approx \frac{h^{(\ell_0)}(0) z^{\ell_0}}{\ell_0!}, \tag{5.49}$$

which, along with (5.30) and (5.38), implies that the scaled abandonment process can be approximated by

$$\breve{A}(t) \approx \frac{n^{\ell_0/2} h^{(\ell_0)}(0)}{\lambda^{\ell_0}(\ell_0+1)!} \int_0^t (\breve{Q}(s))^{\ell_0+1} \, \mathrm{d}s.$$

This approximation implies that the abandonment process depends on the hazard rate function primarily through $h^{(\ell_0)}(0)$, the nonzero derivative at the origin with the lowest order. It also implies that the scaled abandonment rate at time t is approximately

$$\int_0^{\breve{Q}(t)} h\Big(\frac{\sqrt{n}u}{\lambda}\Big) \, \mathrm{d}u \approx \frac{n^{\ell_0/2} h^{(\ell_0)}(0)}{\lambda^{\ell_0}(\ell_0+1)!} (\breve{Q}(t))^{\ell_0+1}. \tag{5.50}$$

In the hazard rate scaling, the scaled queue length in equilibrium q_0 satisfies

$$\lambda = n\mu + \sqrt{n} \int_0^{q_0} h\Big(\frac{\sqrt{n}u}{\lambda}\Big) \, \mathrm{d}u. \tag{5.51}$$

If (5.50) holds, it turns out to be

$$\lambda \approx n\mu + \frac{n^{(\ell_0+1)/2} h^{(\ell_0)}(0)}{\lambda^{\ell_0}(\ell_0+1)!} q_0^{\ell_0+1},$$

which gives us

$$q_0 \approx \frac{1}{\sqrt{n}} \left(\frac{\lambda^{\ell_0}(\ell_0+1)!(\lambda - n\mu)}{h^{(\ell_0)}(0)} \right)^{1/(\ell_0+1)}. \tag{5.52}$$

The scaled queue length process fluctuates around this equilibrium length. Correspondingly, the instantaneous abandonment rate changes around an equilibrium level, too. This observation motivates us to take

$$\alpha = \frac{n^{\ell_0/2} h^{(\ell_0)}(0)}{\lambda^{\ell_0}(\ell_0+1)!} q_0^{\ell_0} \tag{5.53}$$

for the auxiliary $GI/Ph/n+M$ queue. With this setting, the original queue and the auxiliary queue have comparable abandonment rates when the scaled queue length is close to q_0. For any $q_1 > q_0$, when the scaled queue length is q_1 in both queues, the abandonment rate in the auxiliary queue is lower because

$$\alpha q_1 < \frac{n^{\ell_0/2} h^{(\ell_0)}(0)}{\lambda^{\ell_0}(\ell_0+1)!} q_1^{\ell_0+1}.$$

Hence, when the queue length is longer than q_0, it decays slower in the auxiliary queue than in the original queue. Consequently, the decay rate of $\hat{X}$ is slower than that of X and the reference density of $\hat{X}$ can work for this diffusion model.

The above discussion suggests a product reference density in (5.46) with

$$r_j(z) = \begin{cases} \exp\left(-\dfrac{(z+\theta_j\beta_0)^2}{1+c_a^2} \right) & \text{if } z < 0, \\ \exp\left(-\dfrac{\alpha(z-p_jq_0)^2}{\mu(c_a^2+c_s^2)} + \dfrac{\alpha p_j^2 q_0^2}{\mu(c_a^2+c_s^2)} - \dfrac{\theta_j^2\beta_0^2}{1+c_a^2} \right) & \text{if } z \geq 0, \end{cases} \tag{5.54}$$

where q_0 follows (5.52) and α follows (5.53).

The above reference density fails when $\rho = 1$ and $\ell_0 > 0$, because $q_0 = 0$ by (5.52) and thus α is zero in (5.53). In this case, we can still choose a reference density by (5.46) and (5.54) but using a traffic intensity ρ that is slightly larger than one. Because the tail of the queue length becomes heavier as ρ increases, a reference density for model (5.40) with $\rho > 1$ must have a comparable or slower decay rate than the stationary density of the model with $\rho = 1$.

The reference density in (5.46) and (5.54) that exploits the lowest-order nonzero derivative at the origin may fail when the hazard rate function has a rapid change near the origin. In this case, the Taylor's approximation in (5.49) may not be satisfactory when the queue length is not short enough. Such an example is discussed in Section 5.5.4. Choosing a reference density for that is more flexible. In addition, the above procedure cannot choose a reference density when all $h^{(\ell)}(0)$'s are zero, i.e., the hazard rate function is zero in a neighborhood of the origin. This topic will be explored in future research.

5.4.4 Truncation hypercube

Once the reference density has been determined, we can choose the truncation hypercube K in (5.19) by the procedure below. First, pick a small number $\varepsilon_0 > 0$. Then,

choose a hypercube K such that

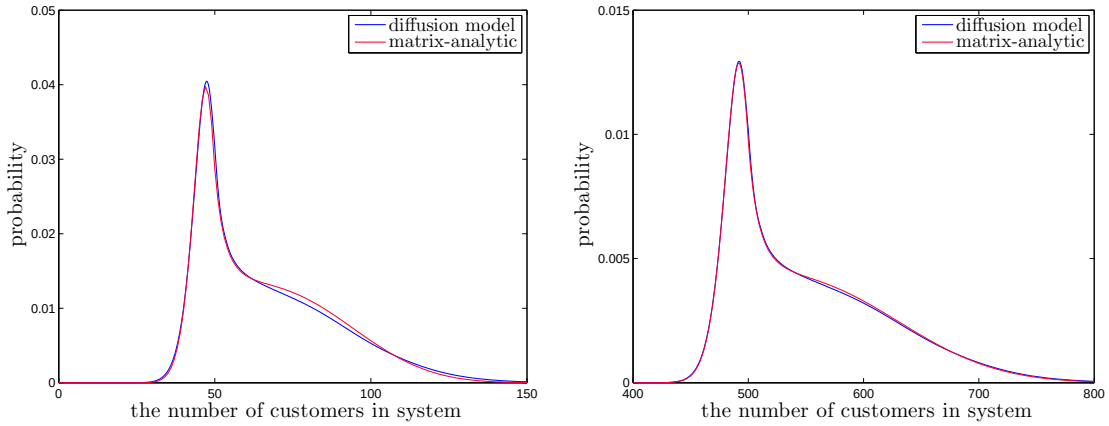
$$\int_{\mathbb{R}^d \setminus K} r(x)\,\mathrm{d}x < \varepsilon_0. \tag{5.55}$$

When ε_0 is small enough, the influence of the reference density outside K is negligible in computing the stationary density.

5.5 Numerical examples

Several numerical examples are presented in this section. In each example, we compute the stationary distribution of the number of customers in a many-server queue using a diffusion model and the proposed algorithm. We assume that the customer arrivals follow a homogeneous Poisson process and the service times follow a two-phase hyperexponential distribution with mean one, i.e., the system is an $M/H_2/n + GI$ queue with $c_a^2 = 1$ and $\mu = 1$. In such a queue, there are two types of customers. One type has a shorter mean service time than the other, and the service times of either type are iid following an exponential distribution. We approximate this queue by a two-dimensional diffusion process X. When the patience time distribution is exponential, both (5.35) and (5.40) yield the same diffusion process. When the patience time distribution is non-exponential, we use model (5.40) as it is more accurate. The results computed using the diffusion models are compared with the ones obtained either by the matrix-analytic method or by simulation. Please refer to [34, 38] for the implementation of the matrix-analytic method. All simulation results are obtained by averaging twenty runs and in each run, the queue is simulated for one hundred thousand time units.

In the proposed algorithm, all numerical integration is implemented by a Gauss–Legendre quadrature rule. See, e.g., [33]. When computing $A_{i\ell}$ or v_i in (5.18), the integrand is evaluated at eight points in each dimension. In the numerical examples,



(a) $\rho = 1.141$ and $n = 50$ (b) $\rho = 1.045$ and $n = 500$

Figure 1: The stationary distribution of the customer number in the $M/H_2/n + M$ queue.

the tail probability

$$\mathbb{P}[X_1(\infty) + X_2(\infty) > z] = \int_{\{x \in \mathbb{R}^2 : x_1 + x_2 > z\}} g(x)\,\mathrm{d}x \quad \text{for some } z \in \mathbb{R} \tag{5.56}$$

is also computed, where $X(\infty) = (X_1(\infty), X_2(\infty))'$ is the two-dimensional random vector having probability density g. The integral in (5.56) is computed by adding up the integrals over the finite elements that intersect with the set $\{x \in \mathbb{R}^2 : x_1 + x_2 > z\}$, and the integral over each finite element is again computed using a Gaussian-Legendre quadrature formula. Because the indicator function has jumps inside certain finite elements, we use sixty-four points in each dimension when evaluating the integrand over each finite element.

5.5.1 Example 1: an $M/H_2/n + M$ queue

Consider an $M/H_2/n + M$ queue that has an exponential patience time distribution. We are interested in such a queue because its customer-count process $N = \{N(t) : t \geq 0\}$ is a quasi-birth-death process. The stationary distribution of that can be computed by the matrix-analytic method.

In this example, we take $\alpha = 0.5$ for the rate of the exponential patience time

Table 1: Performance measures of the $M/H_2/n+M$ queue.

(a) $\rho = 1.141$ and $n = 50$

	Model (5.35)	Matrix-analytic
Mean queue length	17.27	17.16
Abandonment fraction	0.1512	0.1503
$\mathbb{P}[N(\infty) > 45]$	0.8675	0.8523
$\mathbb{P}[N(\infty) > 50]$	0.6785	0.6726
$\mathbb{P}[N(\infty) > 100]$	0.08700	0.07436
$\mathbb{P}[N(\infty) > 130]$	0.008662	0.003299

(b) $\rho = 1.045$ and $n = 500$

	Model (5.35)	Matrix-analytic
Mean queue length	54.17	54.05
Abandonment fraction	0.05181	0.05173
$\mathbb{P}[N(\infty) > 470]$	0.9701	0.9694
$\mathbb{P}[N(\infty) > 500]$	0.6838	0.6818
$\mathbb{P}[N(\infty) > 600]$	0.2244	0.2229
$\mathbb{P}[N(\infty) > 750]$	0.008233	0.006395

distribution and

$$p = (0.9351, 0.0649)' \quad \text{and} \quad \nu = (9.354, 0.072)'$$

for the hyperexponential service time distribution. The mean service time of the second-type customers is more than one hundred times longer than that of the first type. Although over ninety percent of customers are of the first type, the fraction of its workload is merely ten percent, i.e., $\theta = (0.1, 0.9)'$. Such a distribution has a large squared coefficient of variation $c_s^2 = 24$.

The queue is approximated by the two-dimensional piecewise OU process X in (5.35). Because the service time distribution is hyperexponential, P is a zero matrix and thus $R = \mathrm{diag}(\nu)$. By (5.36) and (5.37), the drift coefficient of X is

$$b(x) = \begin{pmatrix} -p_1\mu\beta_0 - \nu_1(x_1 - p_1(x_1+x_2)^+) - p_1\alpha(x_1+x_2)^+ \\ -p_2\mu\beta_0 - \nu_2(x_2 - p_2(x_1+x_2)^+) - p_2\alpha(x_1+x_2)^+ \end{pmatrix} \tag{5.57}$$

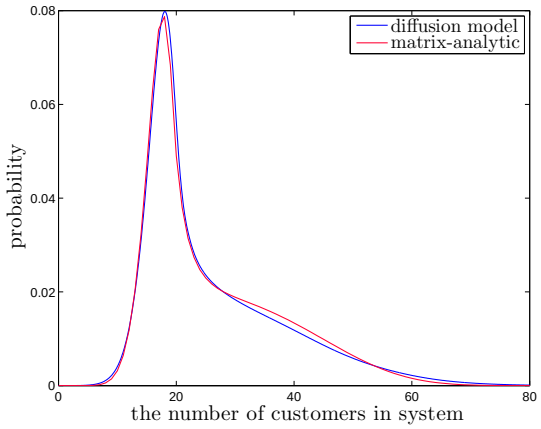


Figure 2: The stationary distribution of the customer number in the $M/H_2/n+M$ queue, with $\rho = 1.112$ and $n = 20$.

and the covariance matrix of the diffusion coefficient is

$$\Sigma(x) = \begin{pmatrix} p_1\mu(\rho + (\rho \wedge 1)) & 0 \\ 0 & p_2\mu(\rho + (\rho \wedge 1)) \end{pmatrix} \tag{5.58}$$

for all $x \in \mathbb{R}^2$.

Three scenarios are considered in this example, in all of which the queue is overloaded. In the first two scenarios, there are $n = 50$ and 500 servers, respectively. The arrival rates are $\lambda = 57.071$ and 522.36, or equivalently, $\rho = 1.141$ and 1.045. By (5.24), $\beta_0 = -1$ in both scenarios. The third scenario, with $n = 20$ servers, will be presented shortly.

To compute the stationary distribution of X, we use a product reference density given by (5.46) and (5.48). To generate basis functions by the finite element method, we set the truncation rectangle $K = [-7, 32] \times [-7, 32]$, which is obtained by (5.55) with $\varepsilon_0 = 10^{-7}$, and use a lattice mesh in which all finite elements are 0.5×0.5 squares.

Once the stationary density of X is obtained, one can approximately produce the distribution of $N(\infty)$, the stationary number of customers in system. Note that the

probability density of $X_1(\infty) + X_2(\infty)$ is given by

$$g_N(z) = \int_{-\infty}^{+\infty} g(x_1, z - x_1)\, \mathrm{d}x_1 \quad \text{for } z \in \mathbb{R}.$$

The distribution of $N(\infty)$ can be approximated by

$$\mathbb{P}[N(\infty) = i] \approx \frac{1}{\sqrt{n}} g_N\Big(\frac{i-n}{\sqrt{n}}\Big) \quad \text{for } i = 0, 1, \ldots$$

For the first two scenarios, the distributions of $N(\infty)$ obtained by the diffusion model are illustrated in Figure 1. In the same figure, the stationary distributions computed by the matrix-analytic method are plotted, too. We see good agreement in Figure 1. Comparing the two scenarios, we also find out that the diffusion model in (5.35) is more accurate when the number of servers n is larger. This observation is consistent with Theorem 4.3.

The matrix-analytic method can be used in this example because the three-dimensional process $\{(Q(t), Z_1(t), Z_2(t)) : t \geq 0\}$ forms a continuous-time Markov chain and the customer-count process N is a quasi-birth-death process. Clearly, $N(t) = Q(t) + Z_1(t) + Z_2(t)$. At time t, N is said to be at level ℓ if $N(t) = \ell$. In this example, level ℓ consists of $\ell + 1$ states if $\ell \leq n$ and it contains $n + 1$ states if $\ell > n$. In the matrix-analytic method, the transition rate matrices between adjacent levels are exploited to compute the stationary distribution of N iteratively. Each iteration requires $O(n^3)$ arithmetic operations. For this queue, the transition rate matrices at different levels are different because the abandonment rate depends on the queue length. For implementation purposes, we assume in the algorithm that at level $\ell > \ell_0$ for some $\ell_0 \gg n$, the abandonment rate at level ℓ is $\alpha(\ell_0 - n)$ rather than $\alpha(\ell - n)$. In other words, the transition rate matrices at level ℓ are invariant with respect to ℓ when $\ell > \ell_0$. We take $\ell_0 = n + 2000$ in all numerical examples. The extra error caused by this modification is negligible, because in this queue, the queue length is in the order of $O(n^{1/2})$ and the chance of the customer number exceeding ℓ_0 is extremely rare.

To investigate the diffusion model in (5.35) quantitatively, we list some steady-state performance measures in Table 1. They include the mean queue length, the fraction of abandoning customers, and the probabilities that the number of customers exceeds certain levels. Using the diffusion model,

$$\text{the mean queue length} \approx \sqrt{n} \int_{\mathbb{R}^2} (x_1 + x_2)^+ g(x)\,\mathrm{d}x$$

and

$$\text{the mean number of idle servers} \approx \sqrt{n} \int_{\mathbb{R}^2} (x_1 + x_2)^- g(x)\,\mathrm{d}x.$$

It follows from the latter approximation that

$$\text{the abandonment fraction} \approx 1 - \frac{\mu}{\lambda}\Big(n - \sqrt{n} \int_{\mathbb{R}^2} (x_1 + x_2)^- g(x)\,\mathrm{d}x\Big).$$

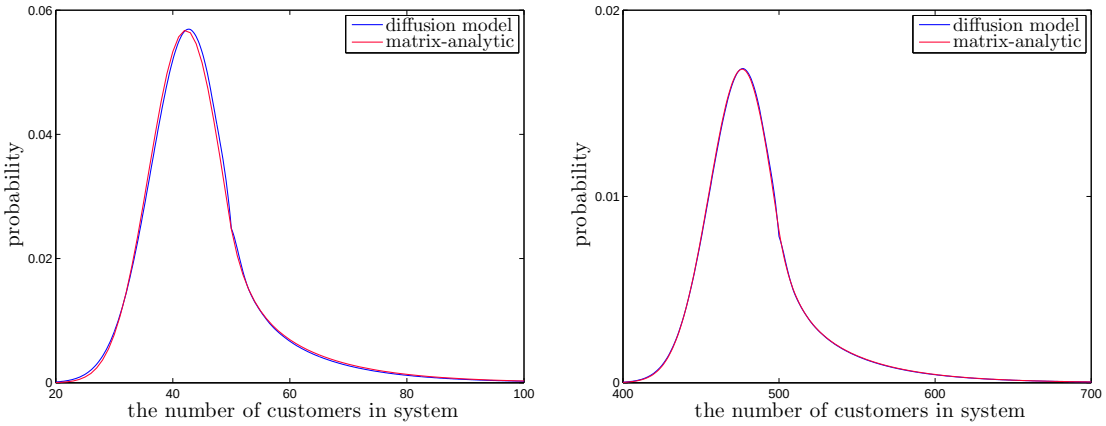
In the table, the tail probability $\mathbb{P}[N(\infty) > \ell]$ is approximated by

$$\mathbb{P}[N(\infty) > \ell] \approx \mathbb{P}\Big[X_1(\infty) + X_2(\infty) > \frac{1}{\sqrt{n}}(\ell - n)\Big] \quad \text{for } \ell = 0, 1, \ldots$$

and $\mathbb{P}[X_1(\infty) + X_2(\infty) > (\ell - n)/\sqrt{n}]$ is computed via (5.56). In both scenarios, the diffusion model produces satisfactory numerical estimates.

The computational complexity of the proposed algorithm, whether in computation time or in memory space, does not change with the number of servers n. In contrast, the matrix-analytic method becomes computationally expensive when n is large. In particular, the memory usage becomes a serious constraint when a huge number of iterations are required. For the $n = 500$ scenario in this example, it took around two hours to finish the matrix-analytic computation and the peak memory usage is nearly five gigabytes. Using the diffusion model and the proposed algorithm, it took less than one minute and the peak memory usage is less than two hundred megabytes on the same computer. See Section 5.6.4 for more discussion on the computational complexity.

Although the diffusion model is motivated and derived from the theory of many-server queues, it is still relevant for a queue with a modest number of servers. In



(a) $\rho = 0.8586$ and $n = 50$ (b) $\rho = 0.9553$ and $n = 500$

Figure 3: The stationary distribution of the customer number in the $M/H_2/n$ queue.

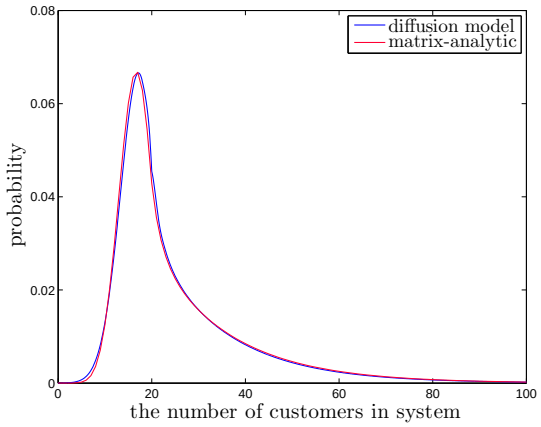


Figure 4: The stationary distribution of the customer number in the $M/H_2/n$ queue, with $\rho = 0.8882$ and $n = 20$.

the third scenario, there are $n = 20$ servers and the arrival rate is $\lambda = 22.24$. Thus, $\rho = 1.112$ and $\beta_0 = -0.5$. In the proposed algorithm, we keep the same truncation rectangle and lattice mesh as in the previous two scenarios, and the reference density is again from (5.46) and (5.48). As illustrated in Figure 2, the diffusion model can still capture the exact stationary distribution for a queue with as few as twenty servers.

Table 2: Performance measures of the $M/H_2/n$ queue.

(a) $\rho = 0.8586$ and $n = 50$

	Model (5.35)	Matrix-analytic
Mean queue length	2.267	2.419
$\mathbb{P}[N(\infty) > 40]$	0.6908	0.6578
$\mathbb{P}[N(\infty) > 50]$	0.2072	0.2012
$\mathbb{P}[N(\infty) > 70]$	0.03395	0.03655
$\mathbb{P}[N(\infty) > 100]$	0.003537	0.003494

(b) $\rho = 0.9553$ and $n = 500$

	Model (5.35)	Matrix-analytic
Mean queue length	8.753	8.800
$\mathbb{P}[N(\infty) > 450]$	0.9038	0.9005
$\mathbb{P}[N(\infty) > 500]$	0.2285	0.2263
$\mathbb{P}[N(\infty) > 600]$	0.01910	0.01908
$\mathbb{P}[N(\infty) > 700]$	0.002241	0.001903

5.5.2 Example 2: an $M/H_2/n$ queue

In this example, an $M/H_2/n$ queue without abandonment is considered. The hyper-exponential service time distribution has

$$p = (0.5915, 0.4085)' \quad \text{and} \quad \nu = (5.917, 0.454)'.$$

Thus, $c_s^2 = 3$ and $\theta = (0.1, 0.9)'$. Since there is no abandonment, we must take $\rho < 1$ in order for the system to reach a steady state.

The diffusion model in (5.35) with $\alpha = 0$ is used. The drift and the diffusion coefficients of X are given by (5.57) and (5.58). The first scenario has $n = 50$ servers and the second scenario has $n = 500$ servers. The respective arrival rates are $\lambda = 42.929$ and 477.64. Hence, $\rho = 0.8586$ and 0.9553, both yielding $\beta_0 = 1$. The product reference density is given by (5.46) and (5.47). With $\varepsilon_0 = 10^{-7}$, the truncation rectangle is set by (5.55) to be $K = [-7, 35] \times [-7, 35]$, which is divided into 0.5×0.5 finite elements.

The stationary distribution of the number of customers in system is shown in Figure 3. In both scenarios, the diffusion model produces a good approximation of

the result by the matrix-analytic method. As in the previous example, the diffusion model is more accurate when the system scale is larger. Several performance measures in steady state are listed in Table 2. As in Table 1, satisfactory agreement can be found between the two approaches.

The third scenario has $n = 20$ servers with arrival rate $\lambda = 17.76$. Then, $\rho = 0.8882$ and $\beta_0 = 0.5$. With $\varepsilon_0 = 10^{-7}$, the truncation rectangle is taken to be $K = [-7, 79] \times [-7, 79]$. The lattice mesh consists of 0.5×0.5 finite elements. The distribution of $N(\infty)$ is shown in Figure 4. For a queue without abandonment, the diffusion model is still useful when the number of servers is modest.

5.5.3 Example 3: an $M/H_2/n + E_k$ queue

Table 3: Performance measures of the $M/H_2/n + E_k$ queue with $\rho < 1$.

(a) $\rho = 0.8586$ and $n = 50$

	$+E_2$		$+E_3$	
	Model (5.40)	Simulation	Model (5.40)	Simulation
Mean queue length	0.9820	1.061	1.201	1.302
Abandonment fraction	0.007974	0.008592	0.005629	0.006115
$\mathbb{P}[N(\infty) > 35]$	0.8881	0.8745	0.8896	0.8762
$\mathbb{P}[N(\infty) > 40]$	0.6755	0.6399	0.6798	0.6448
$\mathbb{P}[N(\infty) > 50]$	0.1671	0.1581	0.1788	0.1707
$\mathbb{P}[N(\infty) > 60]$	0.03238	0.03353	0.04420	0.04584

(b) $\rho = 0.9553$ and $n = 500$

	$+E_2$		$+E_3$	
	Model (5.40)	Simulation	Model (5.40)	Simulation
Mean queue length	4.960	5.048	6.455	6.569
Abandonment fraction	0.001689	0.001729	0.0007611	0.0007931
$\mathbb{P}[N(\infty) > 450]$	0.9003	0.8964	0.9022	0.8984
$\mathbb{P}[N(\infty) > 480]$	0.4759	0.4643	0.4859	0.4746
$\mathbb{P}[N(\infty) > 500]$	0.1995	0.1966	0.2151	0.2124
$\mathbb{P}[N(\infty) > 550]$	0.02798	0.02841	0.04412	0.04458

Consider an $M/H_2/n + E_k$ queue, where $k > 1$ is a positive integer and $+E_k$ signifies an Erlang-k patience time distribution. In this queue, each patience time is the sum of k stages and the stages are iid having an exponential distribution with

Table 4: Performance measures of the $M/H_2/n + E_k$ queue with $\rho > 1$.

(a) $\rho = 1.141$ and $n = 50$

	$+E_2$		$+E_3$	
	Model (5.40)	Simulation	Model (5.40)	Simulation
Mean queue length	15.03	14.94	19.44	19.31
Abandonment fraction	0.1332	0.1334	0.1303	0.1305
$\mathbb{P}[N(\infty) > 45]$	0.9568	0.9490	0.9704	0.9645
$\mathbb{P}[N(\infty) > 50]$	0.8780	0.8648	0.9169	0.9066
$\mathbb{P}[N(\infty) > 70]$	0.3325	0.3121	0.5037	0.4761
$\mathbb{P}[N(\infty) > 90]$	0.008153	0.009354	0.03033	0.03422

(b) $\rho = 1.045$ and $n = 500$

	$+E_2$		$+E_3$	
	Model (5.40)	Simulation	Model (5.40)	Simulation
Mean queue length	76.50	76.20	119.5	119.1
Abandonment fraction	0.04438	0.04437	0.04340	0.04337
$\mathbb{P}[N(\infty) > 480]$	0.9857	0.9846	0.9946	0.9940
$\mathbb{P}[N(\infty) > 500]$	0.9390	0.9363	0.9770	0.9756
$\mathbb{P}[N(\infty) > 600]$	0.3115	0.3051	0.6733	0.6645
$\mathbb{P}[N(\infty) > 700]$	0.0009757	0.0009658	0.04260	0.04358

mean $1/\varpi$. When $k > 1$, the probability density at zero of an Erlang-k distribution is zero. The diffusion model in (5.35) does not have a stationary distribution when the queue is overloaded. Hence, we evaluate the diffusion model in (5.40) that exploits the patience time hazard rate scaling. In the following numerical experiments, we take $k = 2$ or 3 for the Erlang-k distribution and set $\varpi = k$, so the mean patience time is one unit time. The hyperexponential service time distribution is taken to be identical to that in Section 5.5.2.

The hazard rate function of the Erlang-k distribution is

$$h(t) = \frac{\varpi^k t^{k-1}}{(k-1)! \sum_{\ell=0}^{k-1} \frac{\varpi^\ell t^\ell}{\ell!}} \quad \text{for } t \geq 0.$$

For the diffusion model (5.40), it follows from (5.41) that the drift coefficient of X is

$$b(x) = \begin{pmatrix} -p_1\mu\beta_0 - \nu_1(x_1 - p_1(x_1 + x_2)^+) - p_1\eta((x_1 + x_2)^+) \\ -p_2\mu\beta_0 - \nu_2(x_2 - p_2(x_1 + x_2)^+) - p_2\eta((x_1 + x_2)^+) \end{pmatrix} \tag{5.59}$$

where

$$\eta(z) = \int_0^z h\Big(\frac{\sqrt{n}u}{\lambda}\Big)\,\mathrm{d}u = \varpi z - \frac{\lambda}{\sqrt{n}}\log\left(\sum_{m=0}^{k-1}\frac{n^{m/2}\varpi^m z^m}{m!\lambda^m}\right) \quad \text{for } z \geq 0.$$

The first two scenarios has $n = 50$ and 500 servers, respectively. Their respective arrival rates are $\lambda = 42.929$ and 477.64. Hence, $\rho = 0.8586$ and 0.9553, both leading to $\beta_0 = 1$. In the proposed algorithm, the reference density is chosen according to (5.46) and (5.47). The truncation rectangle is taken to be $K = [-7, 35] \times [-7, 35]$ and is divided into 0.5×0.5 finite elements. Some performance estimates can be found in Table 3.

The third and fourth scenarios are for the case $\rho > 1$. They have $n = 50$ and 500 servers, and arrival rates $\lambda = 57.071$ and 522.36, respectively. Then, $\rho = 1.141$ and 1.045, both having $\beta_0 = -1$. For these two scenarios, we adopt the reference density in (5.46) and (5.54). When $k = 2$, each patience time has two stages. The hazard rate function of the patience time distribution has $h(0) = 0$ and $h^{(1)}(0) = \varpi^2$, so $\ell_0 = 1$ in (5.52) and (5.53). Because α in (5.53) depends on n, both the reference density and the truncation rectangle change with n. With $\varepsilon_0 = 10^{-7}$, the truncation rectangle is set to be $K = [-7, 13] \times [-7, 13]$ for $n = 50$ and to be $K = [-7, 16] \times [-7, 16]$ for $n = 500$. When $k = 3$, a patience time consists of three stages. In this case, $h(0) = h^{(1)}(0) = 0$ and $h^{(2)}(0) = 8\varpi^3$, so $\ell_0 = 2$. We set $K = [-7, 11] \times [-7, 11]$ for $n = 50$ and $K = [-7, 15] \times [-7, 15]$ for $n = 500$. All truncation rectangles are partitioned into 0.5×0.5 finite elements. The performance estimates are listed in Table 4.

To evaluate the diffusion model (5.40), we list corresponding simulation estimates of the performance measures in both tables. As in the previous examples, the diffusion model produces adequate performance approximations.

Theoretically, the matrix-analytic method can be used in this example as the customer-count process N is also a quasi-birth-death process. But it is impractical because the computational complexity is too high. Consider the case $k = 2$. Let

Table 5: Performance measures of the $M/H_2/n + H_2$ queue.

(a) $\rho = 1.141$ and $n = 50$

	Model (5.35)	Model (5.40)	Simulation
Mean queue length	0.4709	4.869	4.845
Abandonment fraction	0.1714	0.1504	0.1499
$\mathbb{P}[N(\infty) > 40]$	0.9578	0.9749	0.9728
$\mathbb{P}[N(\infty) > 50]$	0.3158	0.6377	0.6111
$\mathbb{P}[N(\infty) > 60]$	1.044×10^{-7}	0.1895	0.1737
$\mathbb{P}[N(\infty) > 70]$	1.097×10^{-11}	0.02568	0.02142

(b) $\rho = 1.045$ and $n = 500$

	Model (5.35)	Model (5.40)	Simulation
Mean queue length	1.475	6.359	6.413
Abandonment fraction	0.05863	0.05517	0.05512
$\mathbb{P}[N(\infty) > 480]$	0.8663	0.8929	0.8881
$\mathbb{P}[N(\infty) > 500]$	0.3192	0.4822	0.4720
$\mathbb{P}[N(\infty) > 520]$	9.274×10^{-5}	0.1074	0.1050
$\mathbb{P}[N(\infty) > 550]$	-4.488×10^{-9}	0.006616	0.006248

$V_1(t)$ and $V_2(t)$ be the respective numbers of waiting customers whose patience times are in the first and in the second stage at time t. For this $M/H_2/n + E_2$ queue, the four-dimensional process $\{(V_1(t), V_2(t), Z_1(t), Z_2(t)) : t \geq 0\}$ is a continuous-time Markov chain. At level ℓ, there are $\ell+1$ states if $\ell \leq n$ and there are $(n+1)(\ell-n+1)$ states if $\ell > n$. The number of states at level ℓ is formidable when ℓ is large. Even if we may truncate the state space using the technique described in Section 5.5.1, the number of states is still too large to apply the matrix-analytic method. In fact, we are not aware of any other numerical methods other than simulation that can produce approximations in Tables 3 and 4.

5.5.4 Example 4: an $M/H_2/n + H_2$ queue

Let us consider an example in which the patience time hazard rate function changes rapidly near the origin. As pointed out by [45], the performance of such a queue is sensitive to the patience time distribution in a neighborhood of zero. A model that exploits the patience time density at zero solely may not produce adequate performance

estimates. In this example, the patience times follow a two-phase hyperexponential distribution that has

$$\check{p} = (0.9, 0.1)' \quad \text{and} \quad \check{\nu} = (1, 200)'.$$

In other words, there are two types of patience times. Ninety percent of patience times are exponentially distributed with mean one and ten percent are exponentially distributed with mean 0.005. We take the same hyperexponential service time distribution as in Sections 5.5.2 and 5.5.3.

The hazard rate function of the hyperexponential patience time distribution is

$$h(t) = \frac{\check{p}_1\check{\nu}_1 \exp(-\check{\nu}_1 t) + \check{p}_2\check{\nu}_2 \exp(-\check{\nu}_2 t)}{\check{p}_1 \exp(-\check{\nu}_1 t) + \check{p}_2 \exp(-\check{\nu}_2 t)} \quad \text{for } t \geq 0.$$

The drift coefficient of X in (5.40) is also given by (5.59) where

$$\begin{aligned}\eta(z) &= \int_0^z h\Big(\frac{\sqrt{n}u}{\lambda}\Big)\,\mathrm{d}u \\ &= -\frac{\lambda}{\sqrt{n}} \log\Big(\check{p}_1 \exp\Big(-\frac{\sqrt{n}}{\lambda}\check{\nu}_1 z\Big) + \check{p}_2 \exp\Big(-\frac{\sqrt{n}}{\lambda}\check{\nu}_2 z\Big)\Big) \quad \text{for } z \geq 0.\end{aligned}$$

In this example, we have

$$h(0) = \check{p}_1\check{\nu}_1 + \check{p}_2\check{\nu}_2 = 20.9 \quad \text{and} \quad h^{(1)}(0) = -\check{p}_1\check{p}_2(\check{\nu}_1 - \check{\nu}_2)^2 = -3564.1.$$

Thus, $\ell_0 = 0$ and the hazard rate function has a steep slope near the origin. Since the zeroth degree Taylor's approximation in (5.49) may bring on too much error, the reference density exploiting the lowest-order nonzero derivative at the origin could be erroneous.

To choose an appropriate reference density, an auxiliary queue is used again. As in Section 5.4.3, the auxiliary queue is an $M/H_2/n + M$ queue that shares the same arrivals and service times with the $M/H_2/n + H_2$ queue. Let $\alpha > 0$ be the rate of the exponential patience time distribution. We take $\alpha = \check{\nu}_1 \wedge \check{\nu}_2$ so that the patience times in the auxiliary queue are all of the type with the longer mean. If the queue

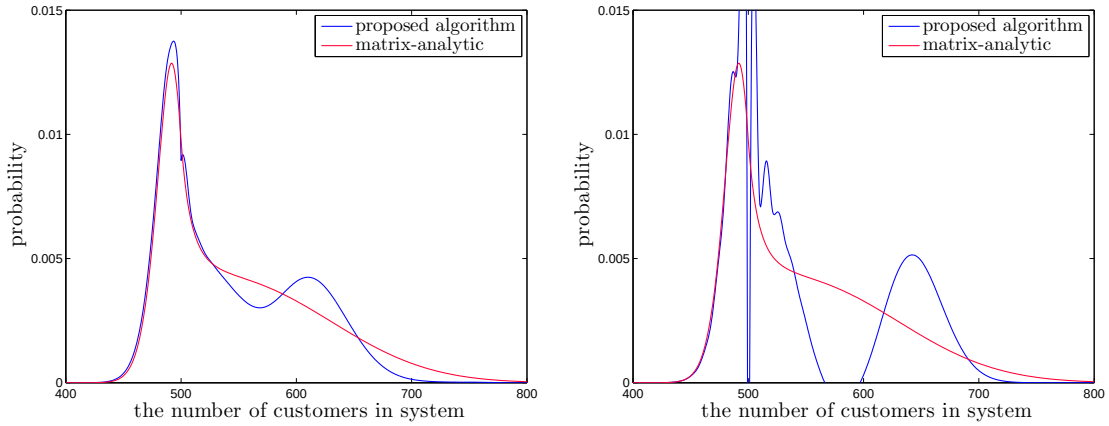
lengths are equal, the abandonment rate in the auxiliary queue must be lower than that in the original queue. Therefore, the queue length decays slower in the former queue and the reference density for model (5.35) of the auxiliary queue should work. This observation leads to a reference density that follows (5.46) and (5.54), but in this example, we take $\alpha = \check{\nu}_1 \wedge \check{\nu}_2$ and solve (5.51) to find q_0.

Two scenarios with $n = 50$ and 500 servers are investigated. The respective arrival rates are $\lambda = 57.071$ and 522.36. Thus, $\rho = 1.141$ and 1.045 and both scenarios have $\beta_0 = -1$. By solving (5.51), we have $q_0 = 0.165$ for the first scenario and $q_0 = 0.0059$ for the second scenario. The reference density follows (5.46) and (5.54) with $\alpha = \check{\nu}_1 = 1$. With $\varepsilon_0 = 10^{-7}$, the truncation rectangle is $K = [-7, 9] \times [-7, 9]$, partitioned into 0.5×0.5 finite elements. The performance estimates obtained by the diffusion model (5.40) are compared with the simulation results in Table 5. The performance estimates are still quite accurate.

We also put the performance estimates produced by the diffusion model (5.35) in this table. For this model, the reference density follows (5.46) and (5.48) with $\alpha = h(0) = 20.9$. In the proposed algorithm, the mesh for model (5.40) is used again. Because in this example, using the patience time density at zero solely cannot capture the behavior of the abandonment process, model (5.35) fails to produce proper performance estimates.

5.6 Implementation issues

The proposed algorithm was implemented using the C++ programming language. The package was tested on both Linux and Windows platforms. In this section, we discuss several important issues arising from the implementation. They are crucial for using the algorithm to solve practical problems. To demonstrate these issues, the second scenario with $n = 500$ servers in Section 5.5.1 is investigated throughout this section. The diffusion model (5.35) is used to approximate the $M/H_2/n + M$ queue.



(a) $K = [-7, 10] \times [-7, 10]$ (b) $K = [-7, 32] \times [-7, 32]$

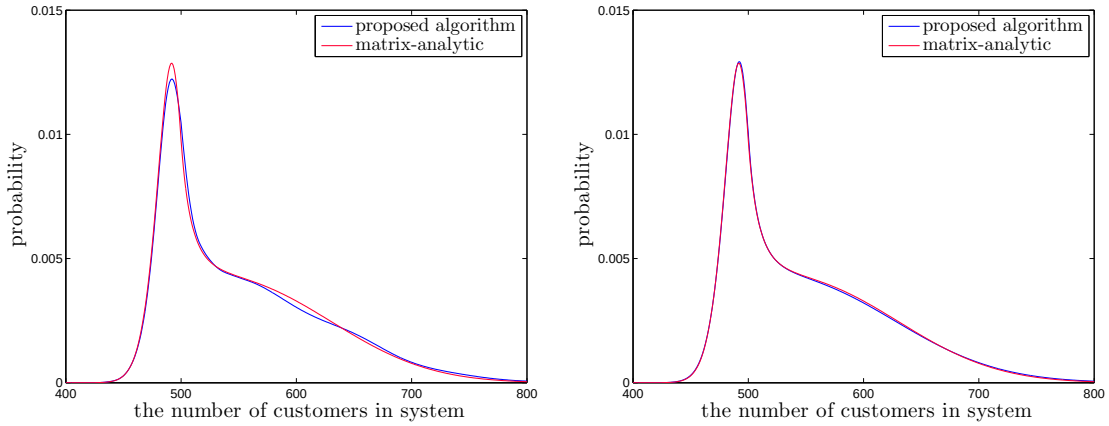
Figure 5: The output of the proposed algorithm with the "naive" reference density.

5.6.1 Influence of the reference density

The reference density plays a key role in the algorithm. If the function r does not satisfy (5.6), the sequence of spaces $\{H_k : k \in \mathbb{N}\}$ may not converge to H in $L^2(\mathbb{R}^d, r)$ and the output of the algorithm may significantly deviate from the exact stationary density. To demonstrate this issue, let us consider a "naive" reference density.

To produce a "naive" reference density, we consider a queue that has the same arrival process and patience time distribution as the $M/H_2/n + M$ queue. This new queue has an exponential service time distribution and its mean service time is equal to that of the $M/H_2/n + M$ queue. For this $M/M/n + M$ queue, the diffusion model (5.35) is a one-dimensional piecewise OU process whose stationary density is given by (5.44). The "naive" reference density is a product reference density in (5.46) with each r_j being the stationary density in (5.44). In other words, the "naive" reference density is obtained by pretending the service time distribution to be exponential.

Let us apply the "naive" reference density. With $\varepsilon_0 = 10^{-7}$, the truncation rectangle is set to be $K = [-7, 10] \times [-7, 10]$ and is partitioned into 0.5×0.5 finite elements. As shown in Figure 5a, the output of the proposed algorithm noticeably deviates from the exact stationary distribution. To further confirm that the "naive" reference



(a) 1.0×1.0 finite elements (b) 0.25×0.25 finite elements

Figure 6: The output of the proposed algorithm with different meshes.

density cannot work, we next test the truncation rectangle $K = [-7, 32] \times [-7, 32]$, which is used in Section 5.5.1 along with the proposed reference density. In this case, the matrix A in (5.17) is close to singular and its condition number is 3.52×10^{190}. Figure 5b manifests the severe error in the algorithm output.

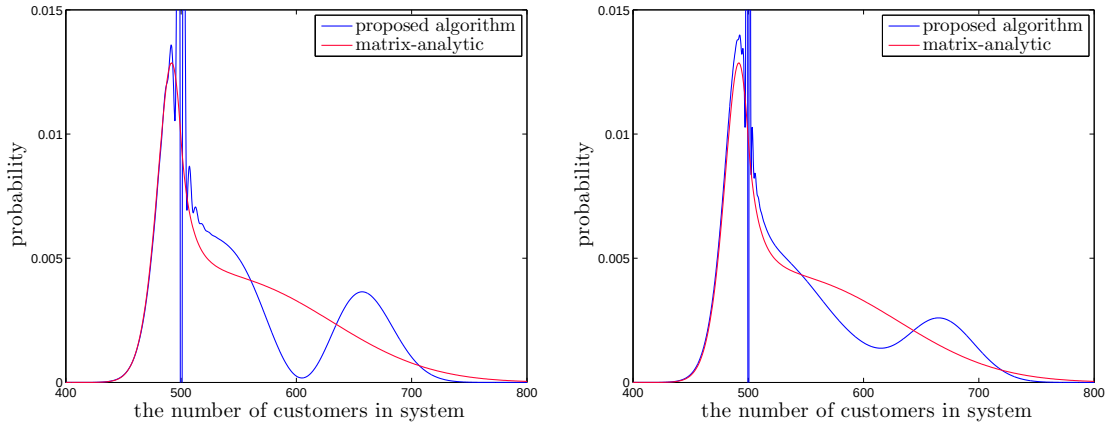
Recall that in this example, the hyperexponential service time distribution has $c_s^2 = 24$. Comparing (5.44) with (5.48), we can tell that the decay rate of the "naive" reference density is much larger than that of the proposed reference density. If Conjecture 5.6 is true, one can expect that the "naive" reference density decays much faster than the stationary density and the second condition in (5.6) may not hold. In this case, the ratio function q is no longer in $L^2(\mathbb{R}^d, r)$ and consequently, the algorithm fails to produce any adequate estimate of the ratio function.

5.6.2 Mesh selection

When both the reference density and the truncation hypercube are fixed, using a finer mesh may produce smaller approximation error. However, a finer mesh yields more basis functions, which in turn lead to a larger condition number for the matrix A in (5.17). If the condition number of A is too large, the round-off error in solving (5.17) becomes considerable. So a finer mesh does not necessarily yield a more accurate

Table 6: The output of the proposed algorithm using different meshes.

	0.5×0.5	0.25×0.25	Matrix-analytic
Mean queue length	54.17	54.17	54.05
Abandonment fraction	0.05181	0.05182	0.05173
$\mathbb{P}[N(\infty) > 470]$	0.9701	0.9702	0.9694
$\mathbb{P}[N(\infty) > 500]$	0.6838	0.6835	0.6818
$\mathbb{P}[N(\infty) > 600]$	0.2244	0.2241	0.2229
$\mathbb{P}[N(\infty) > 750]$	0.008233	0.008246	0.006395



(a) 0.25×0.25 finite elements

(b) 0.125×0.125 finite elements

Figure 7: The output of the proposed algorithm with the "naive" reference density and different meshes.

output.

Let us test different meshes for the second scenario in Section 5.5.1. We keep the same settings for the algorithm except the size of finite elements. The output with 1.0×1.0 finite elements is plotted in Figure 6a. With this mesh, the algorithm does not perform well at the intervals where the stationary density varies quickly. We need a finer mesh to improve the accuracy. In this case, the condition number of A is 5.70×10^{20}. Recall that to produce the curve in Figure 1b, we use a mesh consisting of 0.5×0.5 finite elements. With this mesh, the condition number of A is 1.15×10^{23}. When the element size is further reduced to 0.25×0.25, the condition number of A grows to 7.13×10^{27}. As illustrated in Figure 6b, the output of the algorithm fits the exact stationary distribution well. When we compare Figures 1b and 6b, however,

there is barely any difference noticeable between the algorithm outputs. To confirm that this mesh is not superior to the one with 0.5×0.5 finite elements, we list several performance estimates in Table 6. In this table, the results in Table 1b are duplicated for comparison purposes. The difference between the algorithm outputs using these two meshes is negligible. Considering the modeling error of the diffusion model, we can assert that using 0.5×0.5 finite elements is sufficient to produce an accurate approximation for this queue.

Given an appropriate reference density and the associated truncation hypercube, the above discussion has demonstrated an approach to selecting a mesh. Beginning with two meshes, with one finer than the other, we compare the algorithm outputs using these two meshes. If obvious difference is observed, the coarser mesh should be discarded and a further finer mesh is explored. Continue this procedure until the difference between the outputs of two meshes are negligible. Then, the coarser one of the remaining two is selected as an appropriate mesh.

We would also demonstrate that with an improper reference density, a finer mesh cannot make the algorithm yield an adequate output. Let us go back to the example in Section 5.6.1 with the "naive" reference density. We set the truncation rectangle to be $K = [-7, 32] \times [-7, 32]$ and the size of finite elements to be 0.25×0.25. The output is shown in Figure 7a. Although the curve appears smoother than the one in Figure 5b with 0.5×0.5 finite elements, the output still fails to capture the exact stationary distribution. This time, the condition number of A is 3.91×10^{195}. There is no doubt that such an ill-conditioned matrix will bring about huge round-off error in solving (5.17). A mesh with 0.125×0.125 finite elements is also investigated and the algorithm output is plotted in Figure 7b. The condition number of A increases to 6.35×10^{198} and the algorithm misses the target as well.

Table 7: The output of the proposed algorithm with different quadrature orders.

	$m = 4$	$m = 8$	$m = 16$	Matrix-analytic
Mean queue length	54.17	54.17	54.17	54.05
Abandonment fraction	0.05181	0.05181	0.05181	0.05173
$\mathbb{P}[N(\infty) > 470]$	0.9701	0.9701	0.9701	0.9694
$\mathbb{P}[N(\infty) > 500]$	0.6833	0.6838	0.6839	0.6818
$\mathbb{P}[N(\infty) > 600]$	0.2245	0.2244	0.2244	0.2229
$\mathbb{P}[N(\infty) > 750]$	0.008235	0.008233	0.008232	0.006395

Table 8: Computation time (in seconds) of the proposed algorithm using different meshes.

	1.0×1.0	0.5×0.5	0.25×0.25	0.125×0.125
Dimension m_C	5776	23716	96100	386884
Constructing A and v	6.63	27.3	109	455
Solving (5.17)	0.0780	0.359	2.29	18.2

5.6.3 Gauss–Legendre quadrature

Before solving the linear system (5.17), we must generate the matrix A and the vector v whose entries are given by (5.18). We follow a Gauss–Legendre quadrature rule to compute the integral for each entry. The integral is taken over a two-dimensional rectangle and the quadrature rule evaluates the integrand at m points in each dimension. The results are more accurate when a larger m is used. In Section 5.5, we take $m = 8$ in the numerical examples. Here, we briefly discuss the impact of the order m.

Several performance estimates are listed in Table 7. We keep the same settings for the algorithm except the quadrature order in each dimension. For the convenience of comparison, the results in Table 1b are duplicated in this table. Clearly, the Gauss–Legendre quadrature of order $m \geq 4$ is sufficiently accurate for our purposes.

5.6.4 Computational complexity

Let d, the dimension of the diffusion model, be fixed. The size of A is $m_C \times m_C$ where m_C is the dimension of the functional space C given by (5.23). The matrix A is sparse. There are at most 6^d nonzero entries in each row or column. Hence, it takes

$O(m_C)$ arithmetic operations to construct A. We may apply Gaussian elimination to solve the linear system (5.17). When the basis functions are properly ordered, the nonzero entries of A are confined to a diagonally bordered band of width $O(m_C^{(d-1)/d})$. Hence, solving (5.17) requires $O(m_C^{(2d-1)/d})$ operations as $m_C \to \infty$.

The computation time (measured by seconds) for various meshes can be found in Table 8, where we list both the time for constructing A and v and the time for solving (5.17). When computing A and v, we follow a Gauss–Legendre quadrature rule with $m = 8$ points in each dimension. The truncation rectangle is set to be $K = [-7, 32] \times [-7, 32]$. Each mesh is obtained by setting the size of finite elements. The dimension m_C increases by around four times as the width of each finite element is reduced by half. The proposed algorithm is tested on a laptop with a 2.66GHz Intel Core 2 Duo processor and eight gigabytes memory. Both A and v are produced by our C++ package. The linear system (5.17) is solved by Matlab. These two parts are connected via a MEX interface that comes with Matlab.

CHAPTER VI

FUTURE DIRECTIONS

This thesis focuses on queues with many parallel servers and customer abandonment. Such a queue is used to model a customer call center. We first presented an asymptotic relationship between the abandonment process and the queue length process. Using this relationship, we proved a set of limit theorems for $G/Ph/n+GI$ queues in the QED regime. Motivated by the diffusion limit theorems, we proposed two diffusion models for many-server queues and developed a numerical algorithm for analyzing the diffusion models.

Our future research may include the following topics.

6.1 Distributional insensitivity to service times

The diffusion limit for $G/Ph/n+GI$ queues in Theorem 4.3 depends on the entire service time distribution. Such dependence is in sharp contrast to the diffusion limits for single-server queues that depend on the service time distribution only through its first two moments [31]. Despite the lack of such an invariance principle in many-server queues, Conjecture 5.6 implies that only the first two moments have an influence on the limiting decay rate. In contrast to the Gaussian tail conjectured for $GI/GI/n+GI$ queues, the limiting tail for $GI/GI/n$ queues has an exponential decay rate; see (5.42). For $GI/GI/n$ queues, the limiting decay rate does not depend on the service time distribution beyond the first two moments either. We intend to prove Conjecture 5.6 by constructing two sets of queues with modified service and patience times. They serve as a lower and an upper bound for the original queues. The decay rate in (5.45) may be justified by showing that these two bounds are asymptotically close.

6.2 Measure-valued limits for $G/GI/n+GI$ queues

Kaspi and Ramanan proved a measure-valued limit process for $G/GI/n$ queues in a recent paper [28]. Using the asymptotic relationship (1.2) and the modular approach described in Section 1.2, we intend to generalize their results to $G/GI/n+GI$ queues in the QED regime. We may also generalize the diffusion limit for $GI/M/n+GI$ queues in the patience time hazard rate scaling, which is proved in [45], to the $G/GI/n+GI$ model as well.

In the framework of [45], the patience time distribution of each queue depends on the index n. For the nth queue, let F^n be the patience time distribution and h^n be the hazard rate function of F^n. They assumed that $h^n(t) = h(\sqrt{n}t)$ for each $x \geq 0$ where $h : \mathbb{R}_+ \to \mathbb{R}_+$ is a bounded function. Then,

$$F^n(t) = 1 - \exp\Big(- \int_0^t h(\sqrt{n}u)\,\mathrm{d}u\Big) \quad \text{for } t \geq 0. \tag{6.1}$$

Because the function h on $[0, \varepsilon]$ captures the hazard rate function for the nth system on $[0, n^{-1/2}\varepsilon]$, this scaling magnifies the hazard rate function in a neighborhood of zero. We seek to prove the following relationship between the abandonment processes and the queue length processes.

Conjecture 6.1. *Consider a sequence of $G/G/n+GI$ queues whose queue length processes satisfy (3.8). Assume that the patience time distribution in the nth queue is given by (6.1). Then, for each $T > 0$,*

$$\sup_{0\leq t\leq T}\Big|\tilde{A}^n(t) - \int_0^t \int_0^{\tilde{Q}^n(s)} h(u)\,\mathrm{d}u\,\mathrm{d}s\Big| \to 0 \quad \textit{in probability as } n \to \infty.$$

Using this relationship, we will try to prove a measure-valued limit process for $G/GI/n+GI$ queues in the hazard rate scaling.

6.3 More on the numerical algorithm

Assume that the stationary density g is twice differentiable in $\mathbb{R}^d$ and vanishes at infinity. Using the basic adjoint relationship (5.5) and applying integration by parts

twice, we have

$$\mathcal{G}^* g(x) = 0 \quad \text{for all } x \in \mathbb{R}^d$$

where $\mathcal{G}^*$ is the adjoint operator of the generator $\mathcal{G}$. Fix a finite domain $K \subset \mathbb{R}^d$ large enough. One can solve the stationary density g by the Dirichlet problem

$$\begin{cases} \mathcal{G}^* g(x) = 0 & \text{for } x \text{ in the interior of } K, \\ g(x) = 0 & \text{for } x \text{ on the boundary of } K. \end{cases}$$

Such a Dirichlet problem can be solved via a finite difference algorithm. Alternatively, for each test function f, one may apply integration by parts once to the basic adjoint relationship to obtain an equation that involves the first derivatives of g and the first derivatives of f. From this weak formulation, fixing a large enough finite domain K and assuming that g is zero on the boundary of K, one may apply a standard Galerkin finite element method to compute the stationary density g on K. See, e.g., [32]. Both the finite difference algorithm and the Galerkin method do not use a reference density. A future research topic is to compare the efficiency and accuracy of these two algorithms with the proposed algorithm in Chapter 5.

The dimension of the functional space C in Section 5.2.3 grows exponentially in d, the dimension of the diffusion model. As a consequence, both the computation time and the memory usage increases exponentially in d. When d is not small, the curse of dimensionality is a serious challenge for the proposed algorithm in Chapter 5 as well as any other algorithms. To reduce the dimension of C, one possible approach is to investigate a reference density that potentially shares more common features with the stationary density. Such a reference density may enable us to compute the stationary density with a moderate number of basis functions when d is not small. Another possible direction to reduce the computational complexity of the algorithm is to investigate a low-rank matrix approximation for the linear system (5.17). The technique of random sampling may be explored. See [27] for more details.

APPENDIX A

A CONTINUOUS MAP

Let d be a fixed positive integer. Given functions $h_1 : \mathbb{R}^{d+1} \to \mathbb{R}$, $h_2 : \mathbb{R}^{d+1} \to \mathbb{R}^d$, and $g_0 : \mathbb{R} \to \mathbb{R}^d$, we wish to define a map $\Upsilon : \mathbb{D}^{d+1} \to \mathbb{D}^{d+1}$. For each $w = (u, v) \in \mathbb{D}^{d+1}$ with $u(t) \in \mathbb{R}$ and $v(t) \in \mathbb{R}^d$ for $t \geq 0$, $\Upsilon(w)$ is defined to be any $y = (x, z) \in \mathbb{D}^{d+1}$ with $x(t) \in \mathbb{R}$ and $z(t) \in \mathbb{R}^d$ for $t \geq 0$ that satisfies

$$x(t) = u(t) + \int_0^t h_1(y(s))\,\mathrm{d}s, \tag{A.1}$$

$$z(t) = v(t) + \int_0^t h_2(y(s))\,\mathrm{d}s + g_0(x(t)). \tag{A.2}$$

We assume that h_1, h_2, and g_0 are Lipschitz continuous. For a function $f : \mathbb{R}^\ell \to \mathbb{R}^m$, it is said to be *positively homogeneous* if

$$f(a\phi) = af(\phi) \quad \text{for any } a > 0 \text{ and } \phi \in \mathbb{R}^\ell.$$

Given $\varphi \in \mathbb{D}^\ell$ and $T > 0$, we set $\|\varphi\|_T = \sup_{0 \leq t \leq T} |\varphi(t)|$.

The following lemma establishes the existence and the continuity of the map Υ.

Lemma A.1. *Assume that h_1, h_2, and g_0 are Lipschitz continuous. (a) For each $w = (u, v) \in \mathbb{D}^{d+1}$ with $u(t) \in \mathbb{R}$ and $v(t) \in \mathbb{R}^d$, there exists a unique $y = (x, z) \in \mathbb{D}^{d+1}$ with $x(t) \in \mathbb{R}$ and $z(t) \in \mathbb{R}^d$ that satisfies (A.1) and (A.2). (b) The map Υ: $\mathbb{D}^{d+1} \to \mathbb{D}^{d+1}$ is Lipschitz continuous in the sense that for each $T > 0$, there exists a constant $c_T > 0$ such that*

$$\|\Upsilon(w) - \Upsilon(\check{w})\|_T \leq c_T \|w - \check{w}\|_T \quad \textit{for any } w, \check{w} \in \mathbb{D}^{d+1}.$$

(c) The map Υ is continuous when the domain $\mathbb{D}^{d+1}$ and the range $\mathbb{D}^{d+1}$ are both endowed with the Skorohod J_1-topology. (d) If, in addition, h_1, h_2, and g_0 are assumed

to be positively homogeneous, then the map Υ *is positively homogeneous in the sense that*

$$\Upsilon(aw) = a\Upsilon(w) \quad \textit{for each } a > 0 \textit{ and each } w \in \mathbb{D}^{d+1}.$$

Proof. Assume that h_1, h_2, and g_0 are Lipschitz continuous with Lipschitz constant $\kappa > 0$. Let $w = (u, v) \in \mathbb{D}^{d+1}$ be given. Let $T > 0$ be fixed for the moment. Define $y^0 = w$ and for each $n \in \mathbb{Z}_+$, let $y^{n+1} = (x^{n+1}, z^{n+1})$ be defined via

$$\begin{aligned} x^{n+1}(t) &= u(t) + \int_0^t h_1(y^n(s))\,\mathrm{d}s, \\ z^{n+1}(t) &= v(t) + \int_0^t h_2(y^n(s))\,\mathrm{d}s + g_0(x^{n+1}(t)) \end{aligned}$$

for $t \in [0, T]$. Setting

$$\Delta^n(t) = \|y^{n+1} - y^n\|_t,$$

because

$$\begin{aligned} z^{n+1}(t) - z^n(t) = &\int_0^t \big(h_2(y^n(s)) - h_2(y^{n-1}(s))\big)\,\mathrm{d}s \\ &+ g_0\Big(v(t) + \int_0^t h_1(y^n(s))\,\mathrm{d}s\Big) - g_0\Big(v(t) + \int_0^t h_1(y^{n-1}(s))\,\mathrm{d}s\Big) \end{aligned}$$

for $t \in [0, T]$, one has

$$\Delta^{n+1}(t) \le (\kappa + \kappa^2) \int_0^t \Delta^n(s)\,\mathrm{d}s \quad \text{for } t \in [0, T].$$

Then, by Lemma 11.3 in [35],

$$\Delta^{n+1}(t) \le (\kappa + \kappa^2)\frac{T^n}{n!} \sup_{0 \le s \le t} \Delta^0(s) \quad \text{for } t \in [0, T].$$

Therefore, similar to (11.22) in [35], $\{y^n : n \in \mathbb{N}\}$ is a Cauchy sequence under the uniform norm $\|\cdot\|_T$. Let $\mathbb{D}([0, T], \mathbb{R}^{d+1})$ be the space of functions $f : [0, T] \to \mathbb{R}^{d+1}$ that are right-continuous on $[0, T)$ and have left limits in $(0, T]$. Since the space $(\mathbb{D}([0, T], \mathbb{R}^{d+1}), \|\cdot\|_T)$ is a complete metric space (being a closed subset of the Banach space of bounded functions defined from $[0, T]$ into $\mathbb{R}^{d+1}$ and endowed with

the uniform norm), $\{y^n : n \in \mathbb{N}\}$ has a limit y that is in $\mathbb{D}([0,T],\mathbb{R}^{d+1})$. One can check that y satisfies (A.1) and (A.2) for $t \in [0,T]$. This proves the existence of the map Υ from $\mathbb{D}([0,T],\mathbb{R}^{d+1})$ to $\mathbb{D}([0,T],\mathbb{R}^{d+1})$.

Now we prove that the map from $\mathbb{D}([0,T],\mathbb{R}^{d+1})$ to $\mathbb{D}([0,T],\mathbb{R}^{d+1})$ is Lipschitz continuous with respect to the uniform norm. Assume that $w, \breve{w} \in \mathbb{D}([0,T],\mathbb{R}^{d+1})$. Let $\Upsilon(w)$ be any solution y such that w and y satisfy (A.1) and (A.2) on $[0,T]$. Similarly, let $\Upsilon(\breve{w})$ be any solution associated with $\breve{w}$. Setting $y = (x,z) = \Upsilon(w)$ and $\breve{y} = (\breve{x},\breve{z}) = \Upsilon(\breve{w})$, then for any $t \in [0,T]$,

$$|x(t) - \breve{x}(t)| \le |w(t) - \breve{w}(t)| + \kappa \int_0^t |\Upsilon(w)(s) - \Upsilon(\breve{w})(s)|\,\mathrm{d}s,$$
$$|z(t) - \breve{z}(t)| \le (1+\kappa)|w(t) - \breve{w}(t)| + (\kappa + \kappa^2) \int_0^t |\Upsilon(w)(s) - \Upsilon(\breve{w})(s)|\,\mathrm{d}s.$$

Hence,

$$|\Upsilon(w)(t) - \Upsilon(\breve{w})(t)| \le (1+\kappa)|w(t) - \breve{w}(t)| + (\kappa + \kappa^2) \int_0^t |\Upsilon(w)(s) - \Upsilon(\breve{w})(s)|\,\mathrm{d}s$$

for $t \in [0,T]$. By Corollary 11.2 in [35]

$$\|\Upsilon(w) - \Upsilon(\breve{w})\|_T \le (1+\kappa)\|w - \breve{w}\|_T \exp((\kappa + \kappa^2)T).$$

Hence, Υ is Lipschitz continuous, which implies part (b) of the lemma. The Lipschitz continuity of Υ as a map from $\mathbb{D}([0,T],\mathbb{R}^{d+1})$ to $\mathbb{D}([0,T],\mathbb{R}^{d+1})$ shows that it is well defined on $[0,T]$. Since $T > 0$ is arbitrary, Υ as a map from $\mathbb{D}^{d+1}$ to $\mathbb{D}^{d+1}$ is well defined. This proves part (a) of the lemma.

Next we prove the continuity of Υ provided that $\mathbb{D}^{d+1}$ is endowed with the Skorohod J_1-topology (see, for example, Section 3 of [55]). Consider a sequence $\{w^n : n \in \mathbb{N}\}$ and w in $\mathbb{D}^{d+1}$ such that $w^n \to w$ as $n \to \infty$. Let $y^n = (x^n, z^n) = \Upsilon(w^n)$ and $y = (x,z) = \Upsilon(w)$. Note that since $y \in \mathbb{D}^{d+1}$ there exists $a_T > 0$ such that

$$\|\Upsilon(w)\|_T < a_T. \tag{A.3}$$

Let Λ be the set of strictly increasing functions $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$, $\lim_{t\to\infty} \psi(t) = \infty$, and

$$\gamma_0(\psi) = \sup_{0\le s<t} \left| \log \frac{\psi(t) - \psi(s)}{t - s} \right| < \infty.$$

Since $w^n \to w$ as $n \to \infty$ in the J_1-topology on $\mathbb{D}^{d+1}$, it follows from Proposition 3.5.3 of [14] that there exists a sequence $\{\psi^n : n \in \mathbb{N}\} \subset \Lambda$ such that

$$\lim_{n\to\infty} \gamma_0(\psi^n) = 0 \tag{A.4}$$

and for each $T > 0$

$$\lim_{n\to\infty} \|w^n - w \circ \psi^n\|_T = 0. \tag{A.5}$$

For each $\psi^n \in \Lambda$, $\psi^n(t)$ is Lipschitz continuous in t. Hence, it is differentiable almost everywhere in t with respect to the Lebesgue measure. Furthermore, it follows from (3.5.5) of [14] that when ψ^n is differential at time t, its derivative $\dot{\psi}^n(t)$ satisfies

$$|\dot{\psi}^n(t) - 1| \le \gamma_0(\psi^n). \tag{A.6}$$

Note that, for $i = 1, 2$

$$\int_0^{\psi^n(t)} h_i\left(y(s)\right) \, \mathrm{d}s = \int_0^t h_i\left(y(\psi^n(s))\right) \dot{\psi}^n(s) \, \mathrm{d}s. \tag{A.7}$$

By (A.1) and (A.7)

$$\begin{aligned} x(\psi^n(t)) &= u(\psi^n(t)) + \int_0^{\psi^n(t)} h_1\left(y(s)\right) \, \mathrm{d}s \\ &= u(\psi^n(t)) + \int_0^t h_1\left(y(\psi^n(s))\right) \dot{\psi}^n(s) \, \mathrm{d}s \\ &= u(\psi^n(t)) + \int_0^t h_1\left(y(\psi^n(s))\right) \, ds - \int_0^t h_1\left(y(\psi^n(s))\right) (1 - \dot{\psi}^n(s)) \, \mathrm{d}s. \end{aligned} \tag{A.8}$$

Similarly, by (A.2) and (A.7)

$$\begin{aligned} z(\psi^n(t)) = v(\psi^n(t)) &+ \int_0^t h_2\left(y(\psi^n(s))\right) \, \mathrm{d}s \\ &- \int_0^t h_2\left(y(\psi^n(s))\right) (1 - \dot{\psi}^n(s)) \, \mathrm{d}s + g_0(x(\psi^n(t))). \end{aligned} \tag{A.9}$$

By (A.1) and (A.8),

$$
\begin{aligned}
&|x^n(t) - x(\psi^n(t))| \\
&\le |u^n(t) - u(\psi^n(t))| + \int_0^t |h_1(y^n(s)) - h_1\left(y(\psi^n(s))\right)| \, \mathrm{d}s \\
&\quad + \int_0^t |h_1\left(y(\psi^n(s))\right) - h_1(0)| \, |1 - \dot{\psi}^n(s)| \, \mathrm{d}s + \int_0^t |h_1(0)| \, |1 - \dot{\psi}^n(s)| \, \mathrm{d}s \\
&\le |w^n(t) - w(\psi^n(t))| + \kappa \int_0^t |y^n(s) - y(\psi^n(s))| \, \mathrm{d}s \\
&\quad + \kappa \int_0^t |y(\psi^n(s))| \, |1 - \dot{\psi}^n(s)| \, \mathrm{d}s + |h_1(0)| \int_0^t |1 - \dot{\psi}^n(s)| \, \mathrm{d}s.
\end{aligned} \tag{A.10}
$$

By (A.2), (A.9) and (A.10)

$$
\begin{aligned}
&|z^n(t) - z(\psi^n(t))| \\
&\le |v^n(t) - v(\psi^n(t))| + \int_0^t |h_2(y^n(s)) - h_2\left(y(\psi^n(s))\right)| \, \mathrm{d}s + |g_0(x^n(t)) - g_0(x(\psi^n(t)))| \\
&\quad + \int_0^t |h_2\left(y(\psi^n(s))\right) - h_2(0)| \, |1 - \dot{\psi}^n(s)| \, \mathrm{d}s + \int_0^t |h_2(0)| \, |1 - \dot{\psi}^n(s)| \, \mathrm{d}s \\
&\le |w^n(t) - w(\psi^n(t))| + \kappa \int_0^t |y^n(s) - y(\psi^n(s))| \, \mathrm{d}s + \kappa \, |x^n(t) - x(\psi^n(t))| \\
&\quad + \kappa \int_0^t |y(\psi^n(s))| \, |1 - \dot{\psi}^n(s)| \, \mathrm{d}s + |h_2(0)| \int_0^t |1 - \dot{\psi}^n(s)| \, \mathrm{d}s \\
&\le (1 + \kappa) \, |w^n(t) - w(\psi^n(t))| + (\kappa + \kappa^2) \int_0^t |y^n(s) - y(\psi^n(s))| \, \mathrm{d}s \\
&\quad + (\kappa + \kappa^2) \int_0^t |y(\psi^n(s))| \, |1 - \dot{\psi}^n(s)| \, \mathrm{d}s + (|h_2(0)| + \kappa \, |h_1(0)|) \int_0^t |1 - \dot{\psi}^n(s)| \, \mathrm{d}s.
\end{aligned} \tag{A.11}
$$

Then (A.10) and (A.11) yield

$$
\begin{aligned}
&|\Upsilon(w^n)(t) - \Upsilon(w)(\psi^n(t))| \\
&\le (1 + \kappa) \, |w^n(t) - w(\psi^n(t))| + (\kappa + \kappa^2) \int_0^t |\Upsilon(w^n)(s) \, \mathrm{d}s - \Upsilon(w)(\psi^n(s))| \, \mathrm{d}s \\
&\quad + (\kappa + \kappa^2) \int_0^t |1 - \dot{\psi}^n(s)| \, |\Upsilon(w)(\psi^n(s))| \, \mathrm{d}s \\
&\quad + (|h_2(0)| + \kappa \, |h_1(0)|) \int_0^t |1 - \dot{\psi}^n(s)| \, \mathrm{d}s.
\end{aligned} \tag{A.12}
$$

It follows from (A.3), (A.4), (A.6) and the dominated convergence theorem that

$$\int_0^t |1-\dot{\psi}^n(s)|\,|\Upsilon(w)(\psi^n(s))|\,\mathrm{d}s \to 0 \quad \text{as } n\to\infty. \tag{A.13}$$

Given $\delta > 0$, by (A.4), (A.6) and (A.13), for n large enough

$$(\kappa+\kappa^2)\int_0^T |1-\dot{\psi}^n(s)|\,|\Upsilon(w)(\psi^n(s))|\,\mathrm{d}s + (|h_2(0)| + c\,|h_1(0)|)\int_0^T |1-\dot{\psi}^n(s)|\,\mathrm{d}s < \frac{\delta}{2},$$

and by (A.5)

$$(1+\kappa)\,\|w^n(\cdot) - w(\psi^n(\cdot))\|_T < \frac{\delta}{2}.$$

By Corollary 11.2 in [35] and (A.12),

$$\|\Upsilon(w^n)(\cdot) - \Upsilon(w)(\psi^n(\cdot))\|_T \le \delta \exp((\kappa+\kappa^2)T)$$

for large enough n. Thus, for each $T > 0$,

$$\lim_{n\to\infty} \|\Upsilon(w^n)(\cdot) - \Upsilon(w)(\psi^n(\cdot))\|_T = 0.$$

Hence, $\Upsilon(w^n) \to \Upsilon(w)$ as $n\to\infty$ in $\mathbb{D}^{d+1}$ in the J_1-topology. This implies part (c) of the lemma.

To prove part (d) of the lemma, for $w \in \mathbb{D}^{d+1}$, assume that w and y satisfy (A.1) and (A.2). Then, for $a > 0$, one can check that ay and aw also satisfy (A.1) and (A.2) because of the positive homogeneity of h_1, h_2, and g_0. Therefore, $\Upsilon(aw) = a\Upsilon(w)$. □

APPENDIX B

PROOF OF PROPOSITION 5.4

Recall that K is the compact support of C and the basis functions of C are given by (5.22). We use $C_0^1(K)$ to denote the set of real-valued functions on a neighborhood of K that are continuously differentiable and have compact support in K. Clearly, $C \subset C_0^1(K)$. For any $f, \check{f} \in C_0^1(K)$, we define an inner product by

$$\langle f, \check{f} \rangle_{D(K)} = \sum_{j=1}^{d} \int_K \frac{\partial f(x)}{\partial x_j} \frac{\partial \check{f}(x)}{\partial x_j} \, \mathrm{d}x$$

and let $W_0^{1,2}(K)$ be the closure of $C_0^1(K)$ in the norm induced by this inner product. Then, $W_0^{1,2}(K)$ is a Hilbert space and $C \subset W_0^{1,2}(K)$.

Proof of Proposition 5.4. Since $\mathcal{G}$ is a linear operator, it suffices to show that for any $f_0 \in C$, we must have $f_0 = 0$ if $\mathcal{G} f_0 = 0$ in $L^2(\mathbb{R}^d, r)$. The uniform elliptic operator $\mathcal{G}$ can be written into the divergence form as in (8.1) of [19], i.e.,

$$\mathcal{G} f(x) = \sum_{j=1}^{d} \check{b}_j(x) \frac{\partial f(x)}{\partial x_j} + \frac{1}{2} \sum_{j=1}^{d} \sum_{\ell=1}^{d} \frac{\partial (\Sigma_{j\ell}(x) \partial f(x)/\partial x_j)}{\partial x_\ell}$$

for each $f \in C_b^2(\mathbb{R}^d)$, where

$$\check{b}_j(x) = b_j(x) - \frac{1}{2} \sum_{\ell=1}^{d} \frac{\partial \Sigma_{j\ell}(x)}{\partial x_\ell}.$$

Let $U \subset \mathbb{R}^d$ be a connected open set that is bounded and contains K. Since $r > 0$ and $\mathcal{G} f_0$ is continuous in the interior of each finite element, we must have $\mathcal{G} f_0 = 0$ in K except on the boundaries of certain finite elements where $\mathcal{G} f_0$ is not defined. Hence, $\mathcal{G} f_0 = 0$ in U in the weak sense (see (8.2) of [19]). Note that b, Σ, and the partial derivatives of Σ are all continuous, so both $\check{b}$ and Σ are bounded in U. Because $f_0 \in W_0^{1,2}(K)$, it follows from Corollary 8.2 of [19] that $f_0 = 0$ in K, and thus $f_0 = 0$ in $\mathbb{R}^d$. □

APPENDIX C

PROOF OF PROPOSITION 5.5

Given a compact set $K \subset \mathbb{R}^d$, let $C_b^2(K)$ be the set of real-valued functions on a neighborhood of K that are twice continuously differentiable with bounded first and second derivatives in K. For each $f \in C_b^2(K)$, define a norm $\|\cdot\|_{H^2(K)}$ by

$$\|f\|_{H^2(K)}^2 = \int_K \left(f^2(x) + \max_{j=1,\ldots,d} \Big(\frac{\partial f(x)}{\partial x_j} \Big)^2 + \max_{j,\ell=1,\ldots,d} \Big(\frac{\partial^2 f(x)}{\partial x_j \partial x_\ell} \Big)^2 \right) r(x)\,\mathrm{d}x.$$

Because both b and Σ are bounded in K, there exists $\kappa_0(K) > 0$ such that

$$\int_K (\mathcal{G}f(x))^2 r(x)\,\mathrm{d}x \le \kappa_0(K)\,\|f\|_{H^2(K)}^2 \quad \text{for all } f \in C_b^2(K). \tag{C.1}$$

Let $\bar{C}_b^2(K)$ be the closure of $C_b^2(K)$ in the above norm. A standard procedure can be used to define the first-order and second-order derivatives for $f \in \bar{C}_b^2(K)$. Then, $\mathcal{G}$ can be extended to $\bar{C}_b^2(K)$ and inequality (C.1) holds for all $f \in \bar{C}_b^2(K)$.

Proof of Proposition 5.5. It suffices to prove that for any $f_0 \in C_b^2(\mathbb{R}^d)$, there exists a sequence of functions $\{\varphi_k \in C_k : k \in \mathbb{N}\}$ such that

$$\|\mathcal{G}\varphi_k - \mathcal{G}f_0\| \to 0 \quad \text{as } k \to \infty.$$

Fix $\varepsilon > 0$. Because $K_k \uparrow \mathbb{R}^d$ as $k \to \infty$, by (5.7) and the Cauchy-Schwartz inequality, there exists $a \in \mathbb{N}$ such that

$$\int_{\mathbb{R}^d \setminus K_a} (\mathcal{G}f_0(x))^2 r(x)\,\mathrm{d}x < \frac{\varepsilon^2}{2}. \tag{C.2}$$

Consider the finite hypercube K_a. By (C.1), there exists $\kappa_0(K_a) > 0$ such that

$$\int_{K_a} (\mathcal{G}f(x))^2 r(x)\,\mathrm{d}x \le \kappa_0(K_a)\,\|f\|_{H^2(K_a)}^2 \quad \text{for all } f \in \bar{C}_b^2(K_a). \tag{C.3}$$

A polynomial can be used to approximate f_0 on K_a. By Proposition 7.1 in the appendix of [14], there exists a polynomial f_{p} such that

$$\|f_{\mathrm{p}} - f_0\|_{H^2(K_a)} < \frac{\varepsilon}{2\sqrt{2\kappa_0(K_a)}}.$$

For the lattice mesh Δ_k, let $\Lambda_{a,k}$ be the set of its nodes in the interior of K_a. For any $k \geq a$, let φ_k be a function in C_k such that $\varphi_k(x) = 0$ for all $x \in \mathbb{R}^d \setminus K_a$ and

$$\varphi_k(x) = f_{\mathrm{p}}(x) \quad \text{and} \quad \frac{\partial \varphi_k(x)}{\partial x_j} = \frac{\partial f_{\mathrm{p}}(x)}{\partial x_j}$$

for $j = 1, \ldots, d$ and $x \in \Lambda_{a,k}$. Clearly, $\varphi_k \in \bar{C}_b^2(K_a)$. Because the sequence of meshes $\{\Delta_k : k \in \mathbb{N}\}$ is regularly refined, there exists a constant $\kappa_1 > 0$ such that $\eta_{\Delta_k} < \kappa_1$ for all $k \geq a$. Using the interpolation error estimate in Theorem 6.6 of [39], we have

$$\|\varphi_k - f_{\mathrm{p}}\|_{H^2(K_a)} \leq \kappa_1^2 \kappa_2 \kappa_3 \Big(\int_{\mathbb{R}^d} r(x)\, \mathrm{d}x \Big)^{1/2} |\Delta_k|^2,$$

where $\kappa_2 > 0$ is a constant independent of Δ_k and f_{p}, and

$$\kappa_3 = \sup \left\{ \left| \frac{\partial^4 f_{\mathrm{p}}(x)}{\partial x_1^{m_1} \cdots \partial x_d^{m_d}} \right| : x \in K_a;\, m_1 + \cdots + m_d = 4 \right\} < \infty.$$

Hence, there exists $\delta_0 > 0$ such that

$$\|\varphi_k - f_{\mathrm{p}}\|_{H^2(K_a)} < \frac{\varepsilon}{2\sqrt{2\kappa_0(K_a)}}$$

whenever $|\Delta_k| < \delta_0$. In this case,

$$\|\varphi_k - f_0\|_{H^2(K_a)} \leq \|\varphi_k - f_{\mathrm{p}}\|_{H^2(K_a)} + \|f_{\mathrm{p}} - f_0\|_{H^2(K_a)} < \frac{\varepsilon}{\sqrt{2\kappa_0(K_a)}}.$$

By (C.3),

$$\int_{K_a} (\mathcal{G}\varphi_k(x) - \mathcal{G}f_0(x))^2 r(x)\, \mathrm{d}x \leq \kappa_0(K_a) \|\varphi_k - f_0\|_{H^2(K_a)}^2 < \frac{\varepsilon^2}{2}. \tag{C.4}$$

It follows from (C.2) and (C.4) that $\|\mathcal{G}\varphi_k - \mathcal{G}f_0\| < \varepsilon$ when $k \geq a$ and $|\Delta_k| < \delta_0$. $\square$

REFERENCES

[1] 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.

[2] Baccelli, F., Boyer, P., and Hébuterne, G., "Single-server queues with impatient customers," *Adv. in Appl. Probab.*, vol. 16, no. 4, pp. 887–905, 1984.

[3] Bassamboo, A. and Randhawa, R. S., "On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers," *Oper. Res.*, vol. 58, no. 5, pp. 1398–1413, 2010.

[4] Billingsley, P., *Convergence of Probability Measures*. New York: Wiley, 2nd ed., 1999.

[5] Borovkov, A. A., "On limit laws for service processes in multi-channel systems," *Siberian Math. J.*, vol. 8, pp. 746–763, 1967.

[6] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L., "Statistical analysis of a telephone call center: a queueing-science perspective," *J. Amer. Statist. Assoc.*, vol. 100, no. 469, pp. 36–50, 2005.

[7] Browne, S. and Whitt, W., "Piecewise-linear diffusion processes," in *Advances in Queueing*, pp. 463–480, Boca Raton, FL: CRC, 1995.

[8] Dai, J. G. and Dieker, A. B., "Nonnegativity of solutions to the basic adjoint relationship for some diffusion processes," *Queueing Syst.* To appear.

[9] Dai, J. G. and Harrison, J. M., "Steady-state analysis of RBM in a rectangle: numerical methods and a queueing application," *Ann. Appl. Probab.*, vol. 1, no. 1, pp. 16–35, 1991.

[10] Dai, J. G. and Harrison, J. M., "Reflected Brownian motion in an orthant: numerical methods for steady-state analysis," *Ann. Appl. Probab.*, vol. 2, no. 1, pp. 65–86, 1992.

[11] Dai, J. G., He, S., and Tezcan, T., "Many-server diffusion limits for $G/Ph/n+GI$ queues," *Ann. Appl. Probab.*, vol. 20, no. 5, pp. 1854–1890, 2010.

[12] Dai, J. G. and Tezcan, T., "Optimal control of parallel server systems with many servers in heavy traffic," *Queueing Syst.*, vol. 59, no. 2, pp. 95–134, 2008.

[13] Dieker, A. B. and Gao, X., "Positive recurrence of piecewise Ornstein-Uhlenbeck processes and common quadratic Lyapunov functions." Preprint, 2011.

[14] Ethier, S. N. and Kurtz, T. G., *Markov Processes: Characterization and Convergence*. New York: Wiley, 1986.

[15] Gamarnik, D. and Goldberg, D. A., "Steady-state GI/GI/n queue in the Halfin–Whitt regime." Preprint, 2011.

[16] Gamarnik, D. and Momčilović, P., "Steady-state analysis of a multiserver queue in the Halfin–Whitt regime," *Adv. in Appl. Probab.*, vol. 40, no. 2, pp. 548–577, 2008.

[17] Gans, N., Koole, G., and Mandelbaum, A., "Telephone call centers: tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

[18] Garnett, O., Mandelbaum, A., and Reiman, M., "Designing a call center with impatient customers," *Manufacturing & Service Operations Management*, vol. 4, no. 3, pp. 208–227, 2002.

[19] Gilbarg, D. and Trudinger, N. S., *Elliptic Partial Differential Equations of Second Order*. Berlin: Springer-Verlag, 2001.

[20] Halfin, S. and Whitt, W., "Heavy-traffic limits for queues with many exponential servers," *Oper. Res.*, vol. 29, no. 3, pp. 567–588, 1981.

[21] Harrison, J. M. and Nguyen, V., "The QNET method for two-moment analysis of open queueing networks," *Queueing Syst.*, vol. 6, no. 1, pp. 1–32, 1990.

[22] Iglehart, D. L. and Whitt, W., "Multiple channel queues in heavy traffic. I," *Adv. in Appl. Probab.*, vol. 2, pp. 150–177, 1970.

[23] Iglehart, D. L. and Whitt, W., "Multiple channel queues in heavy traffic. II. Sequences, networks, and batches," *Adv. in Appl. Probab.*, vol. 2, pp. 355–369, 1970.

[24] Jelenković, P., Mandelbaum, A., and Momčilović, P., "Heavy traffic limits for queues with many deterministic servers," *Queueing Syst.*, vol. 47, no. 1-2, pp. 53–69, 2004.

[25] Johnson, D. P., *Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks*. PhD thesis, University of Wisconsin, 1983.

[26] Kang, W. and Ramanan, K., "Fluid limits of many-server queues with reneging," *Ann. Appl. Probab.*, vol. 20, no. 6, pp. 2204–2260, 2010.

[27] Kannan, R. and Vempala, S., "Spectral algorithms," *Found. Trends Theor. Comput. Sci.*, vol. 4, no. 3-4, pp. 157–288, 2008.

[28] Kaspi, H. and Ramanan, K., "SPDE limits of many-server queues." arXiv: 1010.0330, 2010.

[29] Kaspi, H. and Ramanan, K., "Law of large numbers limits for many-server queues," *Ann. Appl. Probab.*, vol. 21, no. 1, pp. 33–114, 2011.

[30] Kiefer, J. and Wolfowitz, J., "On the theory of queues with many servers," *Trans. Amer. Math. Soc.*, vol. 78, pp. 1–18, 1955.

[31] Kingman, J. F. C., "The heavy traffic approximation in the theory of queues. (With discussion)," in *Proc. Sympos. Congestion Theory (Chapel Hill, N.C., 1964)*, pp. 137–169, Chapel Hill, N.C.: Univ. North Carolina Press, 1965.

[32] Kovalov, P., Linetsky, V., and Marcozzi, M., "Pricing multi-asset American options: a finite element method-of-lines with smooth penalty," *J. Sci. Comput.*, vol. 33, no. 3, pp. 209–237, 2007.

[33] Kress, R., *Numerical Analysis*. New York: Springer-Verlag, 1998.

[34] Latouche, G. and Ramaswami, V., *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia, PA: SIAM, 1999.

[35] Mandelbaum, A., Massey, W. A., and Reiman, M. I., "Strong approximations for Markovian service networks," *Queueing Syst.*, vol. 30, no. 1-2, pp. 149–201, 1998.

[36] Mandelbaum, A. and Pats, G., "State-dependent stochastic networks. I. Approximations and applications with continuous diffusion limits," *Ann. Appl. Probab.*, vol. 8, no. 2, pp. 569–646, 1998.

[37] Mandelbaum, A. and Momčilović, P., "Queues with many servers and impatient customers." Preprint, 2009.

[38] Neuts, M. F., *Matrix-Geometric Solutions in Stochastic Models: an Algorithmic Approach*. Baltimore, MD: Johns Hopkins University Press, 1981.

[39] Oden, J. T. and Reddy, J. N., *An Introduction to the Mathematical Theory of Finite Elements*. New York: Wiley, 1976.

[40] Øksendal, B., *Stochastic Differential Equations: an Introduction with Applications*. Berlin: Springer-Verlag, 6th ed., 2003.

[41] Pang, G., Talreja, R., and Whitt, W., "Martingale proofs of many-server heavy-traffic limits for Markovian queues," *Probab. Surv.*, vol. 4, pp. 193–267, 2007.

[42] Protter, P. E., *Stochastic Integration and Differential Equations*. Berlin: Springer-Verlag, 2nd ed., 2005.

[43] Puhalskii, A. A. and Reiman, M. I., "The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime," *Adv. in Appl. Probab.*, vol. 32, no. 2, pp. 564–595, 2000.

[44] Puhalskii, A. A. and Reed, J. E., "On many-server queues in heavy traffic," *Ann. Appl. Probab.*, vol. 20, no. 1, pp. 129–195, 2010.

[45] Reed, J. and Tezcan, T., "Hazard rate scaling for the $GI/M/n+GI$ queue." Preprint, 2009.

[46] Reed, J. E. and Ward, A. R., "Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic," *Math. Oper. Res.*, vol. 33, no. 3, pp. 606–644, 2008.

[47] Reed, J., "The $G/GI/N$ queue in the Halfin-Whitt regime," *Ann. Appl. Probab.*, vol. 19, no. 6, pp. 2211–2269, 2009.

[48] Reiman, M. I., "Open queueing networks in heavy traffic," *Math. Oper. Res.*, vol. 9, no. 3, pp. 441–458, 1984.

[49] Saure, D., Glynn, P., and Zeevi, A., "A linear programming algorithm for computing the stationary distribution of semimartingale reflected Brownian motion." Preprint, 2009.

[50] Shen, X., Chen, H., Dai, J. G., and Dai, W., "The finite element method for computing the stationary distribution of an SRBM in a hypercube with applications to finite buffer queueing networks," *Queueing Syst.*, vol. 42, no. 1, pp. 33–62, 2002.

[51] Stanford, R. E., "Reneging phenomena in single channel queues," *Math. Oper. Res.*, vol. 4, no. 2, pp. 162–178, 1979.

[52] Stone, C., "Limit theorems for random walks, birth and death processes, and diffusion processes," *Illinois J. Math.*, vol. 7, pp. 638–660, 1963.

[53] Tezcan, T., *State Space Collapse in Many Server Diffusion Limits of Parallel Server Systems and Applications*. PhD thesis, Georgia Institute of Technology, 2006.

[54] Tezcan, T. and Dai, J. G., "Dynamic control of N-systems with many servers: asymptotic optimality of a static priority policy in heavy traffic," *Oper. Res.*, vol. 58, no. 1, pp. 94–110, 2010.

[55] Whitt, W., *Stochastic-Process Limits*. New York: Springer-Verlag, 2002.

[56] Whitt, W., "Efficiency-driven heavy-traffic approximations for many-server queues with abandonments," *Management Science*, vol. 50, no. 10, pp. 1449–1461, 2004.

[57] Whitt, W., "Heavy-traffic limits for the $G/H_2^*/n/m$ queue," *Math. Oper. Res.*, vol. 30, no. 1, pp. 1–27, 2005.

[58] Whitt, W., "Fluid models for multiserver queues with abandonments," *Oper. Res.*, vol. 54, no. 1, pp. 37–54, 2006.

[59] Whitt, W., "Proofs of the martingale FCLT," *Probab. Surv.*, vol. 4, pp. 268–302, 2007.

[60] Williams, R. J., "On the approximation of queueing networks in heavy traffic," in *Stochastic Networks: Theory and Applications* (Kelly, F. P., Zachary, S., and Ziedins, I., eds.), Oxford University Press, 1996.

[61] Zeltyn, S. and Mandelbaum, A., "Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue," *Queueing Syst.*, vol. 51, no. 3-4, pp. 361–402, 2005.

[62] Zhang, J., *Limited Processor Sharing Queues and Multi-server Queues*. PhD thesis, Georgia Institute of Technology, 2009.

INDEX