Large Scale Queueing Systems: Asymptotics and Insights

by

David Alan Goldberg

B.S., Columbia University (2006)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author ....
Sloan School of Management
May 20, 2011

Certified by ....
David Gamarnik
Associate Professor of Operations Research
Thesis Supervisor

Accepted by ....
Dimitris Bertsimas
Boeing Professor of Operations Research
Co-director, Operations Research Center

Large Scale Queueing Systems: Asymptotics and Insights

by

David Alan Goldberg

Submitted to the Sloan School of Management
on May 20, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

Parallel server queues are a family of stochastic models useful in a variety of applications, including service systems and telecommunication networks. A particular application that has received considerable attention in recent years is the analysis of call centers. A feature common to these models is the notion of the 'trade-off' between quality and efficiency. It is known that if the underlying system parameters scale together according to a certain 'square-root scaling law', then this trade-off can be precisely quantified, in which case the queue is said to be in the Halfin-Whitt regime.

A common approach to understanding this trade-off involves restricting one's models to have exponentially distributed call lengths, and restricting one's analysis to the steady-state behavior of the system. However, these are considered shortcomings of much work in the area. Although several recent works have moved beyond these assumptions, many open questions remain, especially w.r.t. the interplay between the transient and steady-state properties of the relevant models. These questions are the primary focus of this thesis.

In the first part of this thesis, we prove several results about the rate of convergence to steady-state for the $M/M/n$ queue, i.e. n-server queue with exponentially distributed inter-arrival and processing times, in the Halfin-Whitt regime. We identify the limiting rate of convergence to steady-state, discover an asymptotic phase transition that occurs w.r.t. this rate, and prove explicit bounds on the distance to stationarity. The results of the first part of this thesis represent an important step towards understanding how to incorporate transient effects into the analysis of parallel server queues.

In the second part of this thesis, we prove several results regarding the steady-state $GI/GI/n$ queue, i.e. n-server queue with generally distributed inter-arrival and processing times, in the Halfin-Whitt regime. We first prove that under minor technical conditions, the steady-state number of jobs waiting in queue scales like the square root of the number of servers. We then establish bounds for the large deviations behavior of this model, partially resolving a conjecture made by Gamarnik and Momcilovic in [43]. We also derive bounds for a related process studied by Reed in [91].

We then derive the first qualitative insights into the steady-state probability that an arriving job must wait for service in the Halfin-Whitt regime, for generally distributed processing times. We partially characterize the behavior of this probability when a certain excess parameter B approaches either 0 or ∞. We conclude by studying the large deviations of the number of idle servers, proving that this random variable has a Gaussian-like tail.

We prove our main results by combining tools from the theory of stochastic comparison [99] with the theory of heavy-traffic approximations [113]. We compare the system of interest to a 'modified' queue, in which all servers are kept busy at all times by adding artificial arrivals whenever a server would otherwise go idle, and certain servers can permanently break down. We then analyze the modified system using heavy-traffic approximations. The proven bounds hold for all n, have representations as the suprema of certain natural processes, and may prove useful in a variety of settings. The results of the second part of this thesis enhance our understanding of how parallel server queues behave in heavy traffic, when processing times are generally distributed.

Thesis Supervisor: David Gamarnik
Title: Associate Professor of Operations Research

Acknowledgments

First, I would like to thank my advisor, David Gamarnik, for all his help, support, and guidance over the last five years. His keen eye for interesting research problems, broad interests, and high standards have greatly elevated my ability as a researcher during my time at MIT.

I would also like to thank my other committee members, Dimitris Bertsimas and Kavita Ramanan, not only for their guidance in completing this thesis, but also for everything else they have done for me. Professor Bertsimas has been a mentor to me throughout my Ph.D., w.r.t. matters both academic and personal. His never-ending energy and support for the MIT Operations Research Center is an inspiration to the entire MIT community. Professor Ramanan helped welcome me into the Applied Probability Community, and has been a source of many stimulating discussions and ideas. She has been very giving in many ways, including arranging for me to visit her for several days earlier in my Ph.D.

In addition, I would like to thank the entire Applied Probability Community, for welcoming me with such open arms. I am now very fortunate to count many of you as my personal friends, and look forward to working with you all in the years to come. Although there are too many names to mention here, I would like to especially thank my undergraduate advisor Ward Whitt. He was the main influence that led me to study Operations Research, and in particular Applied Probability. I am forever indebted to him for his mentoring over the last decade. His devotion and passion for research continues to be an inspiration to the entire Applied Probability Community.

Of course, over the years I have had many excellent professors and mentors. This includes all those I took classes with at MIT, as well as Columbia, especially my other undergraduate advisors Cliff Stein and Rocco Servedio. I would also like to thank the many mentors I had even before arriving at Columbia, especially those at the Rutgers

Young Scholars Program in Discrete Math, and the New Jersey Governors School of Science, for helping spark a passion for mathematics within me at a young age.

Many people at MIT also helped me deal with the more practical elements of getting a Ph.D. at MIT, e.g. making sure I had enough money to live, signed up for classes on time, etc. For this, I am indebted to the MIT Operations Research Center's wonderful staff. Speaking of having enough money to live, I also gratefully acknowledge the support of a National Defense Science and Engineering Graduate Fellowship during the first three years of my Ph.D.

My random walk through life has also been greatly enhanced by many wonderful friendships. I would like to thank the entire MIT Operations Research Center for acting as a second family for me here at MIT. Every time I came too close to the breaking point of insanity, there would always be somebody around to grab coffee or dinner with. I would also like to thank my many friends from outside of the ORC. This list includes many that I have met over the last five years, and many that I have known my entire life. I am grateful for all of them.

None of this would have been possible without the love and support of my family, especially my Mom, Dad, and two loving sisters. Through their many sacrifices, I was afforded the very opportunities that ultimately led to my completing this thesis and earning my Ph.D., and I am eternally grateful for this.

Finally, I would like to thank the love of my life, Margaret Frank. During the many late nights I devoted to the work of this thesis, thinking about her and spending time with her was the only thing that kept me sane. She has always been, and continues to be, the primary source of peace and happiness in my life. She makes me more than just a machine for turning coffee into theorems, and it is to her that I dedicate this thesis.

Contents

List of Figures

Chapter 1

Introduction

1.1 Parallel Server Queues

Parallel server queues are a family of stochastic models useful in a variety of applications, including service systems and telecommunication networks. This family of models captures the notion that in such systems, users arrive according to some stochastic arrival process, receive service in parallel according to some stochastic processing mechanism, and ultimately depart the system. Questions of performance of these models has a rich history in the operations research and applied probability literature, under the heading of queueing theory [26],[58], and goes back to the pioneering work of Erlang [39] and Pollaczek [85].

The modeling power of parallel server queues has led to their being used in a great diversity of applications, including call centers [3],[45],[74],[12]; health care [6],[19]; homeland security [108],[46]; transportation [79],[95]; and manufacturing [80],[13]. This diversity of application has led to the creation of a plethora of queueing models, each customized to a particular application. A unifying feature of many of these models is that at their core lie a handful of fundamental models, which are often used as building blocks in more complicated models. Perhaps the most common of these 'building block' models is the First-Come-First-Serve (FCFS) parallel server queue, in which jobs arrive over time according to some stochastic arrival process, and are served in the order in which they arrive, with jobs waiting in a queue until a server becomes available [111].

A particular queueing application that has received considerable attention in recent years is the analysis of call centers [45],[3],[12]. In such systems, the jobs represent incoming calls, and the servers represent call center agents who answer the calls. It has been observed that queueing models suitable for the analysis of call centers have several underlying features, including a large number of servers, a large arrival rate, and a traffic intensity close to unity. Another feature common to these models is the notion of the 'trade-off' between quality and efficiency. Indeed, the managers of call centers must balance the quality of service they provide to callers, e.g. response time, with the cost of providing the chosen level of service, e.g. staffing costs [45],[3],[12].

A common approach to better understanding this quality-efficiency trade-off involves restricting one's models to have exponentially distributed call lengths, and restricting one's analysis by allowing the relevant cost functions to depend only on the steady-state behavior of the system. However, since it is believed that call lengths are actually log-normally distributed [12], and transient effects can be important in staffing a call center [51],[45], these are considered shortcomings of much work in the area. Although several recent works have moved beyond the assumption of Markovian call lengths [87],[60],[75],[43],[91],[64], many open questions remain, especially w.r.t. the interplay between the transient and steady-state properties of the relevant queueing models. These questions are the primary focus of this thesis.

1.2 Model Formulation

In this section we describe the fundamental model studied in this thesis, emphasizing those features most relevant to our results.

FCFS GI/GI/n queue and stability

The main model studied in this thesis is the well-known FCFS parallel server queue with n servers, in which inter-arrival times are drawn independently and identically distributed (i.i.d.), distributed as some random variable (r.v.) A with mean $\mathbb{E}[A]$, and processing times are drawn i.i.d., distributed as some r.v. S with mean $\mathbb{E}[S]$. In the literature, this model is referred to as the FCFS $GI/GI/n$ queue; we refer the reader to [4] for a formal mathematical treatment. Furthermore, if the inter-arrival or processing times are exponentially distributed, the corresponding GI is replaced by an M; e.g. the FCFS $M/GI/n$ queue is the FCFS $GI/GI/n$ queue with exponentially distributed inter-arrival times. We let $Q(t)$ denote the number in system (number in service + number waiting in queue) at time t.

A classical result of Kiefer and Wolfowitz [68] asserts that if the traffic intensity, defined as $\rho \stackrel{\Delta}{=} \frac{\mathbb{E}[S]}{n\mathbb{E}[A]}$, is strictly less than unity, then the FCFS $GI/GI/n$ queue has a meaningful steady-state distribution. In particular, if $\rho < 1$, then under very minor technical conditions on the inter-arrival and processing time r.v.s A and S, the r.v. $Q(t)$ converges in distribution, as $t \to \infty$, to a limiting r.v. $Q(\infty)$, independent of initial conditions; we refer the reader to [4] for details. It is well-known that this steady-state distribution captures the 'long-run' behavior of the underlying system. Note that $\mathbb{P}\big(Q(\infty) \geq n\big)$ represents the probability that an arriving job at some very large time has to wait for service, and we thus refer to $\mathbb{P}\big(Q(\infty) \geq n\big)$ as 'the steady-state probability of delay'.

Quality-efficiency trade-off and the Halfin-Whitt regime

Designers of large parallel server queueing systems must balance the quality of service (QoS) offered to the system users, e.g. the probability of delay, with the cost of providing that service, e.g. the number of servers. In the literature, several strategies have been identified for striking the desired balance. Different strategies result in the queue operating in three fundamentally different regimes as the system grows, where performance metrics behave very differently in each regime [52],[74]. In the quality driven (QD) regime, there is an abundance of servers, and the overwhelming majority of arriving jobs do not have to wait for service. In the efficiency driven (ED) regime, there are barely enough servers, and the overwhelming majority of arriving jobs must wait for service. In the quality and efficiency driven (QED) regime, there are 'a reasonable number' of servers to handle the given load, and some non-trivial fraction, bounded away from both 0 and 1, of arriving jobs must wait for service. It is generally accepted that well-run call centers should operate in the QED regime, and that deciding exactly what constitutes a 'reasonable number' of servers is a challenging optimization problem [45],[10].

Recall that a fundamental feature of call center models is that there are a large number of servers n and a large arrival rate λ. It is sensible to ask, given that n and λ are both large, how should they scale together so that n represents a 'reasonable number' of servers for the arrival rate λ. Although this question has been studied as far back as Erlang [40] and Jagerman [59], the solution was formally given in [52] by Halfin and Whitt, where it was shown that for the case of exponentially distributed processing times, if n scales like $\lambda + B\lambda^{\frac{1}{2}}$ for strictly positive excess parameter B (or equivalently λ scales like $n - Bn^{\frac{1}{2}}$), while the processing time distribution is held fixed, then by letting B vary between 0 and ∞, one can attain any desired probability of delay in the limit as $\lambda, n \rightarrow \infty$. This provides the desired formalization of the quality-efficiency trade-off, parametrized by B.

As empirical studies suggest that call lengths are not exponentially distributed [12], researchers needed a way to extend the model of [52] to more general call length distributions. This was done in a natural manner, see e.g. [91]. Namely, for any fixed inter-arrival distribution A and processing time distribution S, and fixed excess parameter $B > 0$, we define $\lambda_{n,B} \overset{\Delta}{=} n - Bn^{\frac{1}{2}}$, and let Q^n_B denote the FCFS $GI/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as $A\lambda^{-1}_{n,B}$, and processing times drawn i.i.d. distributed as S. Observe that if $\mathbb{E}[A] = \mathbb{E}[S]$, then as in the Markovian model studied in [52], the traffic intensity equals $1 - Bn^{-\frac{1}{2}}$. Q^n_B will be the primary model studied in this thesis. If $\mathbb{E}[A] = \mathbb{E}[S]$ and n is large, then Q^n_B is called the FCFS $GI/GI/n$ queue in the Halfin-Whitt (H-W) regime. Also, we let $Q^n_B(t)$ denote the number in system (number in service + number waiting in queue) in Q^n_B at time t, and $Q^n_B(\infty)$ the corresponding steady-state r.v., when it exists. When the dependence on B is implicit, we will let Q^n, $Q^n(t)$, and $Q^n(\infty)$ denote Q^n_B, $Q^n_B(t)$, and $Q^n_B(\infty)$ respectively.

1.3 Problem Formulation and Literature Review

In this section we formally pose the main problems that we will address in this thesis, and review the relevant literature. We keep our discussion at a high level, saving details for the chapter-specific introductions.

Rate of convergence to steady-state for the M/M/n queue in the Halfin-Whitt regime

Transient effects can be quite important in staffing a call center [51],[45], and thus it is important to understand the error of steady-state approximations for various performance metrics in the H-W regime. Although little is known about the quality of steady-state approximations for the FCFS $GI/GI/n$ queue with general processing time distribution, much more is known when both the inter-arrival and processing time distribution are Markovian. Indeed, the question of how quickly the $M/M/n$ queue approaches stationarity has a rich history in the queueing literature. In [81], Morse derives an explicit solution for the transient $M/M/1$ queue, and discusses implications for the exponential rate of convergence to stationarity. Similar analyses are carried out by Clarke in [25] and Saaty in [94].

Around the same time, both Ledermann and Reuter [72], and Karlin and McGregor [62], worked out powerful and elegant theories that could be used to give the transient distributions for large classes of birth-death processes (b-d-p), including the $M/M/n$ queue, in terms of certain integrals. Karlin and McGregor (K-M) devote an entire paper [63] to the application of their theory to the $M/M/n$ queue, and discuss implications for the rate of convergence to stationarity. These implications were later made rigorous and expanded on in a series of papers [17],[18],[104],[22]. Let $P(t)$ denotes the matrix of transient probabilities for the $M/M/n$ queue, namely $P_{i,j}(t)$ is the probability that there are j jobs in system at time t, if there are i jobs in system at time 0. Let A denote the generator matrix associated with the $M/M/n$ queue, namely the unique rate matrix s.t. $\frac{d}{dt}P(t) = A \cdot P(t)$ [41],[22]. Then it is demonstrated in [17],[18],[104],[22] that $P_{i,j}(t)$ converges exponentially quickly to its steady-state value, at a rate equal to the absolute value of the supremum of the set of strictly negative real eigenvalues of A. This rate is referred to as the spectral gap γ of the associated Markov chain, and we refer the reader to [65] for details, and an excellent survey on transient Markov chains.

It is well-known that for the positive recurrent $M/M/1$ and $M/M/\infty$ queues, γ can be computed explicitly, and has a simple representation in terms of the underlying system parameters [63]. Unfortunately, for the general positive recurrent $M/M/n$

queue, the known characterizations for γ are cumbersome and hard to use [104]. Significant progress towards understanding the spectral gap of the $M/M/n$ queue was made in a series of papers by van Doorn [102],[103],[104],[105]. Van Doorn used the results of K-M and the theory of orthogonal polynomials to give several alternate characterizations and bounds for the spectral gap of a b-d-p, and applied these to the $M/M/n$ queue. He also showed that for each fixed n, there is a phase transition in the nature of the spectral gap of the $M/M/n$ queue as one varies the traffic intensity [102]. Unfortunately, all of the characterizations given by van Doorn, including that of the underlying phase transition, are again fairly complicated, and van Doorn himself comments in [104] that one is generally better off using the approximations that he gives in the same paper. Van Doorn's work was later extended by Kijima in [69], and similar results were achieved by Zeifman using different techniques in [119].

However, it seems that prior to this thesis, these techniques had not been used to analyze the quality of steady-state approximations in the H-W regime. Recently, Leeuwaarden and Knessl studied the rate of convergence to steady-state of a certain related diffusion [70], proving several results analogous to our own. Also, Kang and Ramanan studied the rate of convergence to steady-state of a related fluid limit [61]. However, many questions associated with the rate of convergence to steady-state of the $M/M/n$ queue in the H-W regime were unresolved prior to this thesis.

Explicit bounds on the distance to steady-state for the M/M/n queue in the Halfin-Whitt regime

For applications, it is often desirable to have explicit bounds on the error of the steady-state approximation, as opposed to just an understanding of its behavior up to exponential order. There are several such results in the literature for the $M/M/n$ queue, including the work of Zeifman [119], Chen [23], van Doorn and Zeifman [106], and van Doorn, Zeifman, and Panfilova [107]. Most of these bounds are given in terms of an explicit prefactor attached to an exponentially decaying term. However, these bounds are generally not studied in the H-W regime, and thus may not scale desirably with n in the H-W regime. Indeed, prior to this thesis no explicit bounds on the distance to steady-state for the $M/M/n$ queue were known to perform well in the H-W regime.

Asymptotic scaling of the steady-state GI/GI/n queue in the Halfin-Whitt regime

A first-order consideration when analyzing a queueing model is the question of how the model's performance scales asymptotically with the underlying system parameters. Since it is believed that the call length distributions arising in practice are not Markovian [12], it is important to understand how the FCFS $GI/GI/n$ queue scales in the H-W regime for generally distributed processing times.

In their paper [52], Halfin and Whitt studied Q^n_B in the H-W regime when S is exponentially distributed, i.e. Q^n_B is a $GI/M/n$ queue. They proved that the sequence of processes $\{(Q^n_B(t) - n)n^{-\frac{1}{2}}, n \geq 1\}$, i.e. the diffusion-scaled number of jobs in system, converges in distribution (as a process) to a non-trivial Markovian diffusion, which we call the H-W diffusion, on compact time intervals. They also proved that the sequence of r.v.s $\{(Q^n_B(\infty) - n)n^{-\frac{1}{2}}, n \geq 1\}$, i.e. the diffusion-scaled steady-state number of jobs in system, converges weakly to the mixture of a Gaussian distribution and an exponential distribution, which coincides with the steady-state of the H-W diffusion.

Similar convergence results under the H-W scaling were subsequently obtained for more general multi-server systems. Puhalski and Reiman treated the case of phase-type processing times in [87]. Jelenkovic, Mandelbaum, and Momcilovic treated the case of deterministic processing times in [60]. Mandelbaum and Momcilovic treated the case of processing times with finite support in [75]. Gamarnik and Momcilovic also treated the case of processing times with finite support, albeit from a different perspective, in [43]. Kaspi and Ramanan treated the case of processing times satisfying a mild technical condition in [64], taking a stochastic partial differential equation approach, and showing that the underlying process is an Ito diffusion, in an appropriate sense. The most general known results, essentially covering the case of generally distributed processing times, were proven by Reed in [91], for a class of restrictive initial conditions. Those results were later extended to general initial conditions by Puhalski and Reed in [88].

However, as the theory of weak convergence generally relies heavily on the assumption of compact time intervals, the most general of these results hold only in the transient regime. Indeed, with the exception of [52] (which treats exponentially distributed processing times), [60] (which treats deterministic processing times), and [43] (which treats processing times with finite support), all of the aforementioned results are for the associated sequence of normalized *transient* queue length distributions only, leaving many open questions about the associated *steady-state* queue length distributions. In particular, in [43] it was shown for the case of processing times with finite support that the sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight. Although the authors conjectured that this result should hold for more general processing time distributions, prior to this thesis no further progress on this question had been achieved, leaving open the question of whether $\big(Q^n(\infty) - n\big)^+ = O(n^{\frac{1}{2}})$.

Large deviations for the steady-state GI/GI/n queue in the Halfin-Whitt regime

In many service systems, one is interested in the probability of rare events, which although unlikely, can have serious consequences. The set of associated questions generally falls under the heading of large deviations theory. There is a rich general theory of large deviations in the literature [30]. Although general theorems, such as the celebrated Gartner-Ellis Theorem, provide a framework for proving such results, it is often challenging to derive explicit solutions and insights into any particular stochastic model. There has been much interest in the large deviations of queues [44], and the question has been previously studied in the H-W regime. In particular, in [43], Gamarnik and Momcilovic showed that for the case of processing times with finite support, the sequence of steady-state queue length distributions (normalized by $n^{\frac{1}{2}}$) has a limit whose tail decays exponentially fast. The authors further proved that this exponential rate of decay (i.e. large deviation exponent) is $-2B(c_A^2 + c_S^2)^{-1}$, where c_A^2 and c_S^2 denote the squared coefficient of variation (s.c.v.) of the inter-arrival and processing times, respectively. Similar results are also known to hold for the case of exponentially distributed processing times [52] and deterministic processing times [60].

However, for generally distributed call lengths, the probability that the normalized steady-state number of jobs waiting in queue exceeds some very large value x was not well-understood prior to this thesis, although it had been conjectured by Gamarnik and Momcilovic that the probability of such a rare event should decay exponentially fast at the same rate $-2B(c_A^2 + c_S^2)^{-1}$. Furthermore, even less was known about the large deviations properties of the number of idle servers.

Probability of delay for the steady-state GI/GI/n queue in the Halfin-Whitt regime

An important property of the H-W regime is that the steady-state probability of delay should scale as some non-trivial function of B as $n \to \infty$. This probability often appears in objective functions used to capture the quality-efficiency trade-off in the H-W regime [10],[76], and thus it is important to understand the scaling of this probability for the optimization of such models. The steady-state probability of delay for exponentially distributed processing times was computed explicitly by Halfin and Whitt in [52], and an explicit formula is also known for the case of deterministic processing times [60]. Gamarnik and Momcilovic give an implicit description (in terms of a certain Markov chain) of the steady-state probability of delay for the case of processing times with finite support, and prove that this probability lies strictly in $(0, 1)$. However, it seems that essentially nothing was known about this important quantity for more general processing time distributions prior to this thesis, and very little was known about the qualitative features of this probability beyond the setting of Markovian or deterministic processing times.

1.4 Organization of the Thesis and Main Contributions

Chapter 2: Rate of convergence to steady-state for the M/M/n queue in the Halfin-Whitt regime

In Chapter 2 we prove several results about the exponential rate of convergence to steady-state for the $M/M/n$ queue in the H-W regime. We identify the limiting rate of convergence to steady-state, and discover an asymptotic phase transition that occurs w.r.t. this rate in the H-W regime. In particular, we demonstrate the existence of a constant $B^* \approx 1.85772$ s.t. for $B \in (0, B^*)$, the error in the steady-state approximation converges exponentially fast to 0 at rate $\frac{B^2}{4}$. For $B > B^*$, the error in the steady-state approximation converges exponentially fast to 0 at a different rate, which is the solution to an explicit equation given in terms of special functions. This result may be interpreted as an asymptotic version of the phase transition proven to occur for any fixed n by van Doorn in [102], unifying several earlier characterizations for the spectral gap of the $M/M/n$ queue [63],[102],[104],[119],[69].

Chapter 3: Explicit bounds on the distance to steady-state for the M/M/n queue in the Halfin-Whitt regime

In Chapter 3, we prove explicit bounds on the distance to stationarity for the $M/M/n$ queue in the H-W regime, when $B < B^*$, e.g. characterizing the error in estimating the transient probability of delay by the corresponding steady-state quantity. Our bounds hold for any sufficiently large fixed n, i.e. number of servers, and scale independently of n in the H-W regime. Also, we use our bounds to provide a heuristic rule-of-thumb which could be used to determine the time it takes an overloaded (underloaded) queueing system to return (probabilistically) to the steady-state. In Chapters 2 - 3, we prove our main results by carefully studying the asymptotics of previously known characterizations for the transient $M/M/n$ queue.

The results of Chapters 2 - 3 represent an important step towards understanding how to incorporate transient effects into the analysis of parallel server queues.

Chapter 4: Asymptotic scaling and large deviations for the steady-state GI/GI/n queue in the Halfin-Whitt regime

In Chapter 4, we prove several results regarding the steady-state $GI/GI/n$ queue in the H-W regime. We first prove that under minor technical conditions, the steady-state queue length scales as the square root of the number of servers. More formally, we prove that there exists an a.s. finite r.v. Q^{∞} s.t. for all $x > 0$, $\limsup_{n\rightarrow\infty} \mathbb{P}\big(Q^n(\infty) > n + xn^{\frac{1}{2}}\big) \leq \mathbb{P}\big(Q^{\infty} > x\big)$, i.e. the sequence $\{\big(Q^n(\infty) - n\big)^+ n^{-\frac{1}{2}}, n \geq 1\}$ is tight. We go on to establish bounds for the large deviations behavior of the steady-state $GI/GI/n$ queue in the H-W regime, proving that the tail of the limiting steady-state queue length decays exponentially fast, with exponent less than or equal to $-2B(c_A^2 + c_S^2)^{-1}$. When the arrival process is Poisson, we prove a matching lower bound on the tail of the limiting steady-state queue length. These results partially resolve a conjecture made by Gamarnik and Momcilovic in [43]. We also derive the first non-trivial bounds for a related process studied by Reed in [91]. In particular, in [91], Reed proved that the queue length of the transient $GI/GI/n$ queue converges weakly to a non-trivial process in the H-W regime, under very general assumptions. However, the associated weak limit is only described implicitly, as the solution to a certain stochastic convolution equation (see [91]). We derive the first non-trivial bounds for this weak limit.

Chapter 5: Probability of delay for the steady-state GI/GI/n queue in the Halfin-Whitt regime

In Chapter 5, we derive the first qualitative insights into the steady-state probability of delay in the H-W regime for generally distributed processing times. In particular, we analyze the probability of delay in the H-W regime for the cases $B \to \infty$ and $B \to 0$. We prove that for any fixed distributions A and S, there exist $\epsilon_1, \epsilon_2 > 0$, depending only on A and S, s.t. the limiting steady-state probability of delay is bounded from above by $\exp\big(-\epsilon_1 B^2\big)$ as $B \to \infty$; and the limiting steady-state probability that an arriving job does not have to wait for service, i.e. no delay, is bounded from below by $\epsilon_2 B$ as $B \to 0$. We then revisit the question of large deviations for the steady-state $GI/GI/n$ queue in the H-W regime, but now examine the probability that the steady-state number of idle servers exceeds some large value x. We prove that there exists $\epsilon > 0$, depending only on A, S, and B, s.t. the tail of the limiting steady-state number of idle servers is bounded from below by $\exp\big(-\epsilon x^2\big)$ as $x \to \infty$. These results match known results for the case of Markovian [52] inter-arrival or processing times, and are thus in a sense tight.

We prove our main results by combining tools from the theory of stochastic comparison of queues [99] with the theory of heavy-traffic approximations for queues [113]. In Chapter 4, we compare Q^n to a 'modified' queue, in which all servers are kept busy at all times by adding artificial arrivals whenever a server would otherwise go idle. We then analyze the modified system in the H-W regime using heavy-traffic approximations. In Chapter 5, we compare Q^n to a different 'modified' queue, in which all servers are kept busy on some fixed time interval, at the end of that time interval certain servers break down and cease functioning, and for the remaining time the remaining functional servers are again kept busy. In both cases, the proven bounds are of a structural nature, hold for all n and all times $t \geq 0$, and have intuitive closed-form representations as the suprema of certain natural processes which converge weakly to Gaussian processes in the H-W regime. In both cases, we use special initial conditions to aid in our analysis, since the steady-state distribution does not depend on initial conditions. In particular, we are able to analyze the relevant steady-state distributions without having to analyze the corresponding transient systems under general initial conditions. Furthermore, the results of Chapters 4 - 5 do not follow from naive infinite-server bounds, which either scale incorrectly, or yield inequalities pointing the other direction. Although we ultimately customize these bounds to the H-W regime to prove our main results, we note that our bounds are in no way limited to that regime, and may prove useful in a variety of settings.

The results of Chapters 4 - 5 represent a step towards understanding how to incorporate more general processing time distributions into the analysis of parallel server queues.

1.5 Summary and Open Questions

In this thesis, we prove several results for the performance of the so-called FCFS $GI/GI/n$ queue in the Halfin-Whitt regime. We characterize the error in the steady-state approximation when inter-arrival and processing times are Markovian, an important step towards understanding how to incorporate transient effects into the analysis of parallel server queues. We also prove bounds for the asymptotic behavior of the steady-state queue length, and the probability of certain rare events associated with the steady-state queue length and the probability of delay. Since our bounds hold for a very general class of processing time distributions, these results enhance our understanding of how parallel server queues behave in heavy traffic, when processing times are generally distributed.

This thesis leaves several interesting directions for future research. There are many open questions related to the interaction between weak convergence and convergence to stationarity. Although the results of Chapters 2 - 3 show that one can uniformly bound the rate of convergence to steady-state in the Halfin-Whitt regime for the case of Markovian inter-arrival and processing times, independent of the number of servers, it is open whether such a result holds for more general processing time distributions. A set of related questions has to do with the 'interchange of limits' in the Halfin-Whitt regime. Namely, it is an open question whether or not the sequence $\{n^{-\frac{1}{2}}(Q^n(\infty) - n)^+, n \geq 1\}$ has a unique weak limit. Furthermore, should such a unique weak limit exist, must it coincide with the long-time behavior of the transient weak limit identified by Reed in [91]? Another related interchange question pertains to the fact that many of our large deviations results hold for the limiting diffusion only. In particular, what can be said about the large deviations properties of the pre-limit systems?

Other interesting questions center around the so-called insensitivity phenomenon in queueing systems. In particular, the results of Chapter 4 - 5 can be interpreted as statements about the universality of certain scalings and behaviors. In Chapter 4, we take a step towards proving that the large deviations behavior depends only on the first and second moments of the underlying distributions, and the excess parameter. In Chapter 5, we prove that as one varies the excess parameter, certain fundamental probabilities always scale in the same way, independent of the particulars of the underlying inter-arrival and processing time distributions. Similar phenomena have been observed about queues in the Halfin-Whitt regime by many authors, but the full extent of this insensitivity is not well-understood. Important further steps include generalizing our large deviations lower bounds to non-Poisson arrival processes, and better understanding the steady-state probability of delay.

On a final note, we believe that the bounding methodology introduced in Chapters 4 - 5 may be applicable to a variety of queueing models, and it would be interesting to pursue a research agenda along these lines. For example, perhaps these tools could be used to investigate systems with abandonments in the Halfin-Whitt regime. This setting is practically relevant, since customer abandonments are an important modeling component in the analysis of call centers [3],[12].

Chapter 2

Rate of Convergence to Steady-state for the M/M/n Queue in the Halfin-Whitt Regime

2.1 Introduction and Literature Review

It is well-known that the steady-state behavior of the $M/M/n$ queue in the H-W regime is quite simple in practice [52], while the transient dynamics are more complicated [52], and it is common to use the steady-state approximation to the transient distribution. Thus it is important to understand the quality of the steady-state approximation. The only work along these lines seems to be the recent paper [70], which studies the transform of the H-W diffusion and proves several results analogous to our own for the H-W diffusion. The key difference is that in this paper we study the pre-limit $M/M/n$ queue, not the limiting diffusion. We note that the relevant transform functions were also studied in [5], although in a different context.

The question of how quickly the positive recurrent $M/M/n$ queue approaches stationarity has a rich history in the queueing literature. In [81], Morse derives an explicit solution for the transient $M/M/1$ queue, and discusses implications for the exponential rate of convergence to stationarity. Similar analyses are carried out in [25] and [94]. Around the same time, two different research groups [72], [62] worked out powerful and elegant theories that could be used to give the transient distributions for large classes of birth-death processes (b-d-p). The transient probabilities are expressed as integrals against a spectral measure $\phi(x)$ that is intimately related to the eigenvalues of the generator of the b-d-p. Karlin and McGregor (K-M) devote an entire paper to the application of their theory to the $M/M/n$ queue, in which they comment explicitly on the relationship between the rate of convergence to stationarity and the support of $\phi(x)$ [63]. This relationship was later formalized in a series of papers by other authors [104],[22]. Recall that the spectral gap γ of a b-d-p is the absolute value of the supremum of the set of strictly negative real eigenvalues of the associated generator matrix A, where $\frac{d}{dt}P(t) = A \cdot P(t)$ [41]. If no such eigenvalues exist, we set $\gamma = \infty$. Then it follows from [22] Theorem 5.3, the discussion in the Introduction of [104], and the results of [63] that

Theorem 1. *For any positive recurrent $M/M/n$ or $M/M/\infty$ queue, $\gamma \in (0, \infty)$. For all i and j, $\lim_{t\to\infty} -t^{-1} \log |P_{i,j}(t) - P_j(\infty)| \geq \gamma$. For at least one pair of i, j, $\lim_{t\to\infty} -t^{-1} \log |P_{i,j}(t) - P_j(\infty)| = \gamma$. Furthermore, $\gamma = \inf\{x : x > 0, d\phi(x) > 0\}$.*

We note that γ is also closely related to the singularities of the Laplace transform of $\phi(x)$, and refer the reader to [63] for details. It is well-known that for the positive recurrent $M/M/1$ and $M/M/\infty$ queues, γ can be computed explicitly. In particular, it is proven in [63] that

Theorem 2. *For the positive recurrent $M/M/1$ queue with arrival rate λ and service rate μ, $\gamma = (\lambda^{\frac{1}{2}} - \mu^{\frac{1}{2}})^2$, and the spectral measure $\phi(x)$ consists of a jump at 0, and an absolutely continuous measure on $[(\lambda^{\frac{1}{2}} - \mu^{\frac{1}{2}})^2, (\lambda^{\frac{1}{2}} + \mu^{\frac{1}{2}})^2]$. For the $M/M/\infty$ queue* *with arrival rate λ and service rate μ, $\gamma = \mu$, and the spectral measure $\phi(x)$ consists of a countably infinite number of jumps, at the points $\{k\mu; k \in Z^+\}$.*

Unfortunately, for the general positive recurrent $M/M/n$ queue, the known characterizations for γ involve computing the roots of high-degree polynomials, which may be computationally difficult. This arises from the fact that for the positive recurrent $M/M/n$ queue with arrival rate λ and service rate μ, the spectral measure $\phi(x)$ consists of three parts [63]. The first part is a jump at 0, which corresponds to the steady-state distribution [63]. The second component is an absolutely continuous measure on the interval $[(\lambda^{\frac{1}{2}} - (n\mu)^{\frac{1}{2}})^2, (\lambda^{\frac{1}{2}} + (n\mu)^{\frac{1}{2}})^2]$, whose density is described in [63]. The third component consists of a set of at most n (but possibly zero) jumps, which all exist on $\left(0, (\lambda^{\frac{1}{2}} - (n\mu)^{\frac{1}{2}})^2\right)$ [63]. The complexity of determining γ arises from the difficulty of locating these jumps [104]. In [63], the set of jumps is expressed in terms of the zeros of a certain polynomial equation.

Significant progress towards understanding these jumps was made in a series of papers by van Doorn [102],[103],[104],[105]. Van Doorn used the K-M representation and the theory of orthogonal polynomials to give several alternate characterizations and bounds for the spectral gap of a b-d-p, and applied these to the $M/M/n$ queue. He also showed in [102] that for each fixed n there is a transition in the nature of the spectral measure of the $M/M/n$ queue as one varies the traffic intensity, proving that

Theorem 3. *For all $n \geq 1$, there exists $\rho^*_n \in [0,1)$ such that for any $M/M/n$ queue satisfying $\frac{\lambda}{n\mu} \geq \rho^*_n$, $\gamma = \left(\lambda^{\frac{1}{2}} - (n\mu)^{\frac{1}{2}}\right)^2$; and for any $M/M/n$ queue satisfying $\frac{\lambda}{n\mu} < \rho^*_n$, $\gamma < \left(\lambda^{\frac{1}{2}} - (n\mu)^{\frac{1}{2}}\right)^2$.*

Unfortunately, all of the characterizations (including that of ρ^*_n) given by van Doorn are again stated in terms of the roots of high-degree polynomials, and van Doorn himself comments in [104] that one is generally better off using the approximations that he gives in the same paper. Van Doorn's work was later extended in

[69], and similar results were achieved using different techniques in [119]. It was also shown in [119] that $\rho_n^* \leq (1 - \frac{1}{n})^2$.

In this chapter, we prove several results about the exponential rate of convergence to steady-state for the $M/M/n$ queue in the H-W regime. We identify the limiting rate of convergence to steady-state, and discover an asymptotic phase transition that occurs w.r.t. this rate in the H-W regime. In particular, we demonstrate the existence of a constant $B^* \approx 1.85772$ s.t. for $B \in (0, B^*)$, the error in the steady-state approximation converges exponentially fast to 0 at rate $\frac{B^2}{4}$. For $B > B^*$, the error in the steady-state approximation converges exponentially fast to 0 at a different rate, which is the solution to an explicit equation given in terms of special functions. This result may be interpreted as an asymptotic version of the phase transition proven to occur for any fixed n by van Doorn in [102], unifying several earlier characterizations for the spectral gap of the $M/M/n$ queue [63],[102],[104],[119],[69].

2.1.1 Outline of chapter

The rest of the chapter proceeds as follows. In Section 2.2, we introduce some notation and state our main results. In Section 2.3, we prove a new characterization for γ_n, which is amenable to asymptotic analysis. Sections 2.4 - 2.7 are devoted to studying the asymptotic properties of this characterization. This culminates with Section 2.7, in which we prove our main results. In Section 2.8 we summarize our main results and present ideas for future research. We include a technical appendix in Section 2.9.

2.2 Main Results

2.2.1 Definitions and notations

We now define several important quantities for the $M/M/n$ queue $\mathcal{Q}^n$, namely the $M/M/n$ queue with arrival rate $\lambda_n = n - Bn^{\frac{1}{2}}$ and service rate $\mu = 1$, where we assume throughout that n is sufficiently large to ensure that $n > \lambda_n + 1$. Recall that $Q^n(t)$ denotes the number in system at time t in $\mathcal{Q}^n$. We define $P^n_{i,j}(t) \triangleq \Pr(Q^n(t) = j | Q^n(0) = i)$, $P^n_j(\infty) \triangleq \Pr(Q^n(\infty) = j)$, $P^n_{i,\leq j}(t) \triangleq \sum_{k=0}^{j} P^n_{i,k}(t)$, and $P^n_{\leq j}(\infty) \triangleq \sum_{k=0}^{j} P^n_k(\infty)$. Let γ_n denote the spectral gap of the associated Markov chain. For a complex-valued function $f(x)$, we let $Z(f)\big(Z^+(f)\big)$ denote the infimum of the set of (strictly positive) real zeros of $f(x)$. Set $Z(f)\big(Z^+(f)\big) = \infty$ if $f(x)$ has no (strictly positive) real zeros. We let $H_k \triangleq \sum_{j=1}^{k} \frac{1}{j}$ denote the kth harmonic number. All logarithms will be base e. Unless otherwise stated, all functions are defined only for real values of x. All empty products are assumed to be equal to unity, and all empty summations are assumed to be equal to zero.

2.2.2 The parabolic cylinder functions

We now briefly review the two-parameter function commonly referred to as the parabolic cylinder function $D_x(z)$, since we will need these functions for the statement (and proofs) of our results. For excellent references on these functions, see [50] Section 8.31 and Section 9.24, [14] Sections 3.3-3.5, and [38] Chapter 8. Let $\Gamma(x)$ denote the Gamma function (see [54], Chapter 8.8). It is stated in [14] that for real x and z, $D_x(z) \in \mathbb{R}$, and

$$D_x(z) = \begin{cases} (\frac{2}{\pi})^{\frac{1}{2}} e^{\frac{z^2}{4}} \int_0^\infty e^{-\frac{t^2}{2}} cos(\frac{\pi}{2}x - zt) t^x dt & \text{if } x \geq 0; \\ \frac{e^{-\frac{z^2}{4}}}{\Gamma(-x)} \int_0^\infty e^{-\frac{t^2}{2} - zt} t^{-(x+1)} dt & \text{if } x < 0. \end{cases} \tag{2.1}$$

$D_x(z)$ takes on a simpler form for $x \in Z$. In particular, it is stated in [50] that for $z \in \mathbb{R}$,

$$D_{-1}(z) = 2^{\frac{1}{2}} e^{\frac{z^2}{4}} \int_{2^{-\frac{1}{2}}z}^{\infty} e^{-t^2} dt, \quad D_0(z) = e^{-\frac{z^2}{4}}, \quad \text{and} \quad D_1(z) = z e^{-\frac{z^2}{4}}. \tag{2.2}$$

Note that since $\Gamma(-x) \in (0, \infty)$ for $x < 0$, (2.1) and (2.2) imply that

$$\text{for all} \quad z \in \mathbb{R} \quad \text{and} \quad x \leq 0, \quad D_x(z) > 0. \tag{2.3}$$

The parabolic cylinder functions arise in several contexts associated with the limits of queueing models, such as the Ornstein-Uhlenbeck limit of the appropriately scaled infinite server queue [56] and various limits associated with the Erlang loss model [116]. We note that the parabolic cylinder functions have been studied as the limits of certain polynomials under the H-W scaling, using tools from the theory of differential equations [32],[33],[37],[89].

2.2.3 Main results

We now state our main results. We begin by describing the asymptotic phase transition that occurs w.r.t. the spectral gap γ_n of the $M/M/n$ queue in the H-W regime. Let

$$v(x, y) \triangleq \begin{cases} \frac{D_x(y)}{D_{x-1}(y)} & \text{if } D_{x-1}(y) \neq 0; \\ \infty & \text{otherwise;} \end{cases}$$

Also, let $z_\infty(x) \triangleq v(x, -B) + B$, $\varphi(B) \triangleq v(\frac{B^2}{4}, -B)$, $\zeta(B) \triangleq \varphi(B) + \frac{B}{2}$, and

$$\Psi_\infty(x) \triangleq \begin{cases} v(x, -B) + \frac{1}{2}(B + (B^2 - 4x)^{\frac{1}{2}}) & \text{if } D_{x-1}(-B) \neq 0, x \leq \frac{B^2}{4}; \\ \infty & \text{otherwise.} \end{cases}$$

Note that $\zeta(B) = \Psi_\infty(\frac{B^2}{4})$. We include a plot of $\zeta(B)$. Let $B^* \triangleq Z^+(\zeta)$. Then

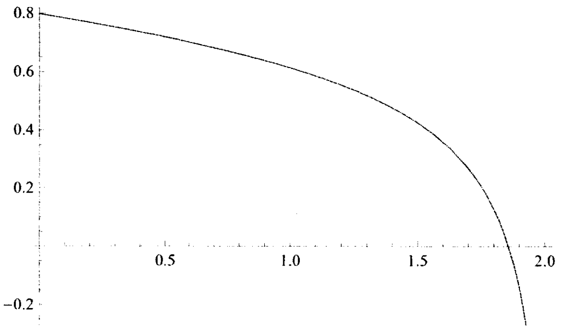

Figure 2-1: $\zeta(B)$

Proposition 1. *$B^* \in [2^{\frac{1}{2}}, 2)$, and $Z^+(\Psi_\infty) < \min(1, \frac{B^2}{4})$ for $B > B^*$. Numerically, $B^* \approx 1.85772$.*

Our main result is that

Theorem 4. *The limit $\gamma_B \triangleq \lim_{n\to\infty} \gamma_n$ exists for all $B > 0$. For $0 < B \leq B^*$, $\gamma_B = \frac{B^2}{4}$. For $B \geq B^*$, $\gamma_B = Z^+(\Psi_\infty)$.*

We include a plot of γ_B.

We note that due to the non-linear manner in which the steady-state probability of delay scales in the H-W regime, the case $0 < B < B^*$ actually encompasses most

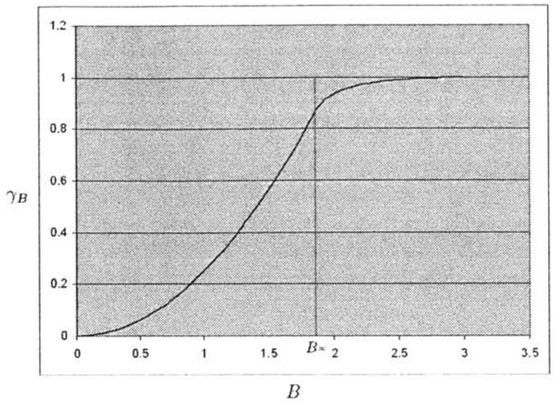

Figure 2-2: γ_B

scenarios of interest. Indeed, it is proven in [52] that

$$\lim_{n\to\infty} P^n_{\leq n}(\infty) = 1 - \big(1 + B\exp(\frac{1}{2}B^2)\int_{-\infty}^{B}\exp(-\frac{1}{2}z^2)dz\big)^{-1}.$$

As this limit is monotone in B, the case $0 < B < B^*$ includes all scenarios for which the steady-state probability of delay is at least .04.

The following corollary may be interpreted as an asymptotic version of Theorem 3, unifying several earlier characterizations for the spectral gap of the $M/M/n$ queue [63],[102],
[104],[119],[69].

Corollary 1. *The ρ^*_n parameter of Theorem 3 satisfies*

$$\lim_{n\to\infty} n^{\frac{1}{2}}(1-\rho^*_n) = B^*.$$

We now give an interpretation of Theorem 4 and Corollary 1. The $M/M/n$ queue behaves like an $M/M/1$ queue when all servers are busy, and an $M/M/\infty$ queue when at least one server is idle. The phase transition of Theorem 4 formalizes this relationship in a new way. For $0 < B < B^*$, the K-M spectral measure of the $M/M/n$ queue in the H-W regime has no jumps away from the origin, and has spectral gap equal to $(\lambda_n^{\frac{1}{2}} - n^{\frac{1}{2}})^2$, two properties shared by the associated $M/M/1$ queue (see Theorem 2). For $B > B^*$, the K-M spectral measure has at least one jump away from the origin, like the associated $M/M/\infty$ queue (whose spectral measure has only jumps and spectral gap equal to 1, see Theorem 2). Another interpretation is that the $M/M/n$ queue cannot approach stationarity faster than either component system would on its own.

2.3 Characterization for γ_n

In this section we give a new characterization for γ_n, which will be amenable to asymptotic analysis.

We begin by associating several functions to the $M/M/n$ queue, as in [69] and [102]. Let

$$f_{n,k}(x) \triangleq \begin{cases} 1 & \text{if k = 0;} \\ 1 + \lambda_n - x & \text{if k = 1;} \\ (\lambda_n + k - x) f_{n,k-1}(x) - \lambda_n (k-1) f_{n,k-2}(x) & \text{otherwise.} \end{cases}$$

By a simple induction argument, for $0 \leq k \leq n$,

$$f_{n,k}(x) = \sum_{j=0}^{k} \binom{k}{j} \lambda_n^j \prod_{i=1}^{k-j} (i - x). \tag{2.4}$$

Also note that

$$Z(f_{n,n-1}) > 1, \quad \text{and} \quad f_{n,n-1}(x) > 0 \quad \text{for} \quad x < Z(f_{n,n-1}). \tag{2.5}$$

Indeed, for $x \leq 1$ and $0 \leq j \leq n-1$, $\prod_{i=1}^{k-j}(i-x) \geq 0$, and by (2.4), $f_{n,n-1}(x)$ is a non-negative sum of such terms with the $(n-1)$st term (λ_n^{n-1}) strictly positive. Let

$$a_n(x) \triangleq \begin{cases} \frac{1}{2}\Big(\lambda_n + n - x - \big((\lambda_n + n - x)^2 - 4\lambda_n n\big)^{\frac{1}{2}}\Big) & \text{if } x \leq (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2; \\ \infty & \text{otherwise.} \end{cases}$$

Note that $a_n(x)$ is real-valued for $x \leq (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$, and therefore

$$a_n(x) = \frac{1}{2}\Big(\lambda_n + n - x - \big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 - x\big)^{\frac{1}{2}}\big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - x\big)^{\frac{1}{2}}\Big). \tag{2.6}$$

Also,

$$a_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) = (\lambda_n n)^{\frac{1}{2}}. \tag{2.7}$$

We also define

$$a_\infty(x) \triangleq \begin{cases} \frac{1}{2}\big(B - (B^2 - 4x)^{\frac{1}{2}}\big) & \text{if } x \leq \frac{B^2}{4}; \\ \infty & \text{otherwise.} \end{cases}$$

Let $\sigma_n(y) \triangleq f_{n,n}(y) - (\lambda_n n)^{\frac{1}{2}} f_{n,n-1}(y)$, $\psi_n(y) \triangleq f_{n,n}(y) - a_n(y) f_{n,n-1}(y)$,

$$z_{n,k}(y) \triangleq \begin{cases} \frac{f_{n,k}(y)}{f_{n,k-1}(y)} & \text{if } f_{n,k-1}(y) \neq 0; \\ \infty & \text{otherwise;} \end{cases}$$

and $z_n(y) \triangleq z_{n,n}(y)$. Let

$$\Psi_n(y) \triangleq \begin{cases} z_n(y) - a_n(y) & \text{if } z_n(y) \neq \infty \text{ or } a_n(y) \neq \infty. \\ \infty & \text{otherwise;} \end{cases}$$

Before proving our new characterization for γ_n, we cite some properties of $z_n(x), a_n(x)$, and $\Psi_n(x)$, as stated in [69], which will be necessary for later proofs.

Lemma 1. *(i) $Z(f_{n,k-1}) \geq Z(f_{n,k})$ for $k \leq n$.*

(ii) For $x \in \big(-\infty, Z(f_{n,n-1})\big)$ and $k \leq n$, $z_{n,k}(x)$ is a strictly positive, continuous, and strictly decreasing function of x.

(iii) For $x \in \big(-\infty, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big]$, $a_n(x)$ is a strictly positive, continuous, strictly increasing function of x.

(iv) For $x \in \Big(-\infty, \min\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2, Z(f_{n,n-1})\big)\Big)$, $\Psi_n(x)$ is a continuous, strictly decreasing function of x. Also, if $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < Z(f_{n,n-1})$, then $\Psi_n(x)$ is continuous at $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$.

We now prove the main result of this section, namely a new characterization for γ_n, which is more amenable to asymptotic analysis. In particular, we prove that

Lemma 2. *(i) If $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 1$ and $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$, then $Z(\Psi_n) = Z^+(\Psi_n)$ is the unique zero of $\Psi_n(x)$ in the interval $\big(0, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)$, and $\gamma_n = Z^+(\Psi_n)$.*

(ii) If $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 1$ and $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) \geq 0$, then $\gamma_n = (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$.

(iii) If $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 \geq 1$,then $Z(\Psi_n) = Z^+(\Psi_n)$ is the unique zero of $\Psi_n(x)$ in the interval $\big(0, 1\big)$, and $\gamma_n = Z^+(\Psi_n)$.

The proof of Lemma 2 relies heavily on a known characterization for γ_n, proven in [69]. Namely,

Theorem 5. *If $Z(\sigma_n) \geq (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$, then $\gamma_n = (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. If $Z(\sigma_n) < (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$, then $\gamma_n = Z(\psi_n)$.*

With Theorem 5 in hand, we now complete the proof of Lemma 2.

Proof of Lemma 2. We begin by proving some properties of Ψ_n and σ_n. From (2.4),

$$\begin{aligned}\Psi_n(0) &= \frac{\sum_{k=0}^{n}\binom{n}{k}\lambda_n^k\prod_{i=1}^{n-k} i}{\sum_{k=0}^{n-1}\binom{n-1}{k}\lambda_n^k\prod_{i=1}^{n-1-k} i} - \frac{1}{2}\Big(\lambda_n + n - \big((\lambda_n+n)^2 - 4\lambda_n n\big)^{\frac{1}{2}}\Big)\\ &= n\frac{\sum_{k=0}^{n}\frac{\lambda_n^k}{k!}}{\sum_{k=0}^{n-1}\frac{\lambda_n^k}{k!}} - \lambda_n \quad > \quad 0.\end{aligned}$$

If $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 \geq 1$, then

$$\begin{aligned}\Psi_n(1) &= \frac{\sum_{k=0}^{n}\binom{n}{k}\lambda_n^k\prod_{i=1}^{n-k}(i-1)}{\sum_{k=0}^{n-1}\binom{n-1}{k}\lambda_n^k\prod_{i=1}^{n-1-k}(i-1)} - \frac{1}{2}\Big(\lambda_n + n - 1 - \big((\lambda_n+n-1)^2 - 4\lambda_n n\big)^{\frac{1}{2}}\Big)\\ &= \frac{\lambda_n^n}{\lambda_n^{n-1}} - \frac{1}{2}\Big(\lambda_n + n - 1 - \big((\lambda_n+n-1)^2 - 4\lambda_n n\big)^{\frac{1}{2}}\Big)\\ &= \frac{1}{2}\Big(\lambda_n - n + 1 + \big((\lambda_n - n + 1)^2 - 4\lambda_n\big)^{\frac{1}{2}}\Big) \quad \leq \quad 0,\end{aligned}$$

since $\Psi_n(1) \in \mathbb{R}$ by (2.5) and Lemma 1.iv, and thus $0 \leq (\lambda_n - n + 1)^2 - 4\lambda_n \leq (n - \lambda_n - 1)^2$. Also,

$$\begin{aligned}\sigma_n(0) &= \sum_{k=0}^{n}\binom{n}{k}\lambda_n^k\prod_{i=1}^{n-k} i - (\lambda_n n)^{\frac{1}{2}}\sum_{k=0}^{n-1}\binom{n-1}{k}\lambda_n^k\prod_{i=1}^{n-1-k} i\\ &= (n-1)!(n\sum_{k=0}^{n}\frac{\lambda_n^k}{k!} - (\lambda_n n)^{\frac{1}{2}}\sum_{k=0}^{n-1}\frac{\lambda_n^k}{k!})\\ &\geq \quad (n-1)!\sum_{k=0}^{n}\frac{\lambda_n^k}{k!}(n - (\lambda_n n)^{\frac{1}{2}}) \quad > \quad 0,\end{aligned}$$

and

$$\begin{aligned}\sigma_n(1) &= \sum_{k=0}^{n}\binom{n}{k}\lambda_n^k\prod_{i=1}^{n-k}(i-1) - (\lambda_n n)^{\frac{1}{2}}\sum_{k=0}^{n-1}\binom{n-1}{k}\lambda_n^k\prod_{i=1}^{n-1-k}(i-1)\\ &= \lambda_n^n - (\lambda_n n)^{\frac{1}{2}}\lambda_n^{n-1} \quad < \quad 0.\end{aligned}$$

We first prove assertion i. Since $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 1$, we have by (2.5) that $\min\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2, Z(f_{n,n-1})\big) = (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$, and $f_{n,n-1}(x) > 0$ on $\big(-\infty, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big]$. Thus from definitions and dividing through by $f_{n,n-1}(y)$, we find that $\sigma_n(y)$ is the same sign as $z_n(y) - (\lambda_n n)^{\frac{1}{2}}$, and that $\psi_n(y)$ is the same sign as $\Psi_n(y)$, on $\big(-\infty, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big]$. Also, by Lemma 1.ii, $z_n(y)$ is strictly positive, continuous, and strictly decreasing on $\big(-\infty, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big]$. Thus by the Intermediate Value Theorem, $\sigma_n(y)$ has a zero on $\big(-\infty, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)$ iff $z_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) - (\lambda_n n)^{\frac{1}{2}} < 0$. By definitions and (2.7), we thus have that $Z(\sigma_n) < (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$ iff $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$. Since by assumption $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$, it will be the case that $Z(\sigma_n) < (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. Furthermore, it follows from the fact that $\Psi_n(0) > 0$ and the continuity and monotonicity of $\Psi_n(y)$ guaranteed by Lemma 1.iv that $Z(\Psi_n)$ will be the unique zero of Ψ_n in the interval $\big(0, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)$. Since $\psi_n(y)$ is the same sign as $\Psi_n(y)$ on $\big(-\infty, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big]$, we will also have that $Z(\psi_n) = Z(\Psi_n)$. This, combined with Theorem 5, completes the proof of assertion i. The proof of assertion ii. proceeds nearly identically to the proof of assertion i. Indeed, the only difference is that in this case by assumption $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) \geq 0$, and thus one concludes that $Z(\sigma_n) \geq (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$, which combined with Theorem 5 proves assertion ii.

We now prove assertion iii. By assumption, $(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2 \geq 1$. Thus since $Z(f_{n,n-1}) > 1$ by (2.5), we have that $f_{n,n-1}(x) > 0$ on $(-\infty, 1]$. Thus from definitions, and dividing through by $f_{n,n-1}(y)$, we find that $\psi_n(y)$ is the same sign as $\Psi_n(y)$ on $(-\infty, 1]$. Since $\sigma_n(y)$ is a polynomial and thus continuous, by the Intermediate Value Theorem and the fact that $\sigma_n(0) > 0$ and $\sigma_n(1) < 0$, we have $Z(\sigma_n) < 1$. Thus $Z(\sigma_n) < (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. By the continuity and monotonicity of $\Psi_n(y)$ guaranteed by Lemma 1.iv and the fact that $\Psi_n(0) > 0$, $\Psi_n(1) < 0$, we will have that $Z(\Psi_n)$ is the unique zero of Ψ_n in the interval $(0,1)$. Furthermore, since $\psi_n(y)$ is the same sign as $\Psi_n(y)$ on $(-\infty, 1]$, we will have that $Z(\psi_n) = Z(\Psi_n)$. This, combined with Theorem 5, completing the proof of assertion iii. □

2.4 Asymptotic Analysis of the Function $\Psi_n(x)$

In this section we derive the asymptotics of $\Psi_n(x)$ and $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)$. In particular, we prove

Theorem 6. *For all $B > 0$ and $x \in (0,1), x \leq \frac{B^2}{4}$, $\lim_{n\to\infty} \lambda_n^{-\frac{1}{2}} \Psi_n(x) = \Psi_\infty(x)$.*

We also prove

Corollary 2. *For $0 < B < 2$, $\lim_{n\to\infty} \lambda_n^{-\frac{1}{2}} \Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) = \zeta(B)$.*

We proceed by separately analyzing the asymptotics of $\big(a_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$ and $\big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$, and then use the fact that by definition $\Psi_n(x)\lambda_n^{-\frac{1}{2}} = \big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}} - \big(a_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$.

We first analyze $a_n(x)$. We begin by proving some bounds for $n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}}$, namely

Lemma 3. $\frac{B}{2} \leq n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}} \leq \frac{B}{2} + O(n^{-\frac{1}{2}})$, *and* $\lim_{n\to\infty}(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}}) = \frac{B}{2}$.

Proof. Note that

$$n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}} = \frac{(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})(n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})}{n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}}} = \frac{B}{2} + \frac{B^2 n^{\frac{1}{2}}}{2(n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2},$$

from which the lemma follows. □

We now study the asymptotics of $a_n(x)$, proving that

Lemma 4. *For* $0 \leq x \leq \frac{B^2}{4}$, $\lim_{n\to\infty} \big(a_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}} = a_\infty(x)$.

Proof. From (2.6) and Lemma 3,

$$\big(a_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}} = \left(Bn^{\frac{1}{2}} - x - \big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - x\big)^{\frac{1}{2}}\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 - x\big)^{\frac{1}{2}}\right)(2\lambda_n^{\frac{1}{2}})^{-1}. \tag{2.8}$$

The lemma then follows from the fact that $\lim_{n\to\infty}(Bn^{\frac{1}{2}} - x)(2\lambda_n^{\frac{1}{2}})^{-1} = \frac{B}{2}$, $\lim_{n\to\infty} \big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - x\big)^{\frac{1}{2}}\big(2\lambda_n^{\frac{1}{2}}\big)^{-1} = 1$, and Lemma 3. □

We now analyze the asymptotics of $z_n(x)$, proving that

Proposition 2. *For* $x \in \big(0, 1\big)$, $\lim_{n\to\infty} \big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}} = z_\infty(x)$.

Suppose that $x \in \big(0, 1\big)$ is fixed. Our proof will use a truncation argument to ensure boundedness of certain quantities, and thus let us fix some integer $T \in [3, \infty)$, and define $C_{T,x} \stackrel{\Delta}{=} \max\big(\frac{2(T+1)T^{-x}}{\prod_{j=1}^{T}(1-\frac{x}{j})}, \frac{(T+1)^2 T^{-x}}{\prod_{j=1}^{T}(1-\frac{x}{j})}\big)$. Then we begin by proving that

Lemma 5. *For all sufficiently large* n, $\big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$ *is at least*

$$\frac{e^{-2T^{-1}}\lambda_n^{\frac{x-1}{2}}\sum_{k=0}^{n-(T+1)}(n-k)^{1-x}e^{-\lambda_n}\frac{\lambda_n^k}{k!}}{e^{2T^{-1}}\lambda_n^{\frac{x-1}{2}}\sum_{k=0}^{\lceil n-T^{-1}n^{\frac{1}{2}}\rceil}k\lambda_n^{-\frac{1}{2}}(n-k)^{-x}\frac{e^{-\lambda_n}\lambda_n^k}{k!} + \frac{2e^{2T^{-1}}T^{-(1-x)}(\frac{n}{\lambda_n})^{\frac{1-x}{2}}}{1-x} + \lambda_n^{\frac{x-1}{2}}C_{T,x}}, \tag{2.9}$$

and at most

$$\frac{e^{2T^{-1}}\lambda_n^{\frac{x-1}{2}}\sum_{k=0}^{n-(T+1)}(n-k)^{1-x}e^{-\lambda_n}\frac{\lambda_n^k}{k!} + \lambda_n^{\frac{x-1}{2}}C_{T,x}n^{-\frac{1}{2}}}{e^{-2T^{-1}}\lambda_n^{\frac{x-1}{2}}\sum_{k=0}^{\lceil n-T^{-1}n^{\frac{1}{2}}\rceil}k\lambda_n^{-\frac{1}{2}}(n-k)^{-x}\frac{e^{-\lambda_n}\lambda_n^k}{k!}}. \tag{2.10}$$

Proof. The proof is deferred to the appendix. □

We now relate the summations appearing in (2.9) and (2.10) to the expectations of certain functions of a Poisson r.v., and then show that these expectations converge to certain integrals as $n \to \infty$. Let X_n denote a Poisson r.v. with mean λ_n, and $Z_n \triangleq (X_n - \lambda_n)\lambda_n^{-\frac{1}{2}}$. For an event $\{E\}$, we let $I(\{E\})$ denote the indicator function of $\{E\}$. Thus

$$X_n = \lambda_n^{\frac{1}{2}} Z_n + \lambda_n, \quad \text{and} \quad n - X_n = Bn^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}} Z_n. \tag{2.11}$$

Furthermore, we define

$$Y_{1,n} \triangleq \big(B(\frac{n}{\lambda_n})^{\frac{1}{2}} - Z_n\big)^{1-x} I\Big(Z_n \leq B(\frac{n}{\lambda_n})^{\frac{1}{2}} - (T+1)\lambda_n^{-\frac{1}{2}}\Big),$$

and

$$Y_{2,n} \triangleq \big(B(\frac{n}{\lambda_n})^{\frac{1}{2}} - Z_n\big)^{-x} I\Big(Z_n \leq (B - T^{-1})(\frac{n}{\lambda_n})^{\frac{1}{2}} + \big(\lceil n - T^{-1}n^{\frac{1}{2}}\rceil - (n - T^{-1}n^{\frac{1}{2}})\big)\lambda_n^{-\frac{1}{2}}\Big).$$

We then have that

Lemma 6. *If* $\lim_{n\to\infty} \mathbb{E}[Y_{1,n}]$, $\lim_{n\to\infty} \mathbb{E}[Y_{2,n}]$, *and* $\lim_{n\to\infty} \mathbb{E}[Z_n Y_{2,n}]$ *all exist, and are finite and strictly positive, then*

$$\liminf_{n\to\infty} \big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}} \geq e^{-4T^{-1}} \frac{\lim_{n\to\infty} \mathbb{E}[Y_{1,n}]}{\lim_{n\to\infty} \mathbb{E}[Y_{2,n}] + \frac{2T^{-(1-x)}}{1-x}},$$

and

$$\limsup_{n\to\infty} \big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}} \leq e^{4T^{-1}} \frac{\lim_{n\to\infty} \mathbb{E}[Y_{1,n}]}{\lim_{n\to\infty} \mathbb{E}[Y_{2,n}]}.$$

Proof. Suppose that $\lim_{n\to\infty} \mathbb{E}[Y_{1,n}]$, $\lim_{n\to\infty} \mathbb{E}[Y_{2,n}]$, and $\lim_{n\to\infty} \mathbb{E}[Z_n Y_{2,n}]$ all exist, and are finite and strictly positive. We begin by expressing the summation appearing in the numerator of (2.9) and (2.10) as an expectation. Namely, $\lambda_n^{\frac{x-1}{2}} \sum_{k=0}^{n-(T+1)} (n - k)^{1-x} e^{-\lambda_n} \frac{\lambda_n^k}{k!}$ is equal to

$$\lambda_n^{\frac{x-1}{2}} \mathbb{E}\Big[\big(Bn^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}} Z_n\big)^{1-x} I\Big(\lambda_n^{\frac{1}{2}} Z_n + \lambda_n \leq n - (T+1)\Big)\Big] \quad = \quad \mathbb{E}[Y_{1,n}] \text{ by (2.11).} \tag{2.12}$$

We now analyze the summation appearing in the denominator of (2.9) and (2.10). Note that $\lambda_n^{\frac{x-1}{2}} \sum_{k=0}^{\lceil n-T^{-1}n^{\frac{1}{2}}\rceil} k(n-k)^{-x}\lambda_n^{-\frac{1}{2}}\frac{e^{-\lambda_n}\lambda_n^k}{k!}$ equals

$$\lambda_n^{\frac{x-1}{2}}\mathbb{E}\left[(Z_n+\lambda_n^{\frac{1}{2}})(Bn^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}}Z_n)^{-x}I\left(\lambda_n^{\frac{1}{2}}Z_n+\lambda_n \le \lceil n-T^{-1}n^{\frac{1}{2}}\rceil\right)\right] \text{ by (2.11).} \quad (2.13)$$

Furthermore, $I\big(Z_n \le (\lceil n-T^{-1}n^{\frac{1}{2}}\rceil - \lambda_n)\lambda_n^{-\frac{1}{2}}\big)$ equals

$$I\left(Z_n \le (B-T^{-1})(\frac{n}{\lambda_n})^{\frac{1}{2}} + \big(\lceil n-T^{-1}n^{\frac{1}{2}}\rceil - (n-T^{-1}n^{\frac{1}{2}})\big)\lambda_n^{-\frac{1}{2}}\right). \quad (2.14)$$

Plugging (2.14) back into (2.13), we find that the summation appearing in the numerator of (2.9) and (2.10) equals $\lambda_n^{-\frac{1}{2}}\mathbb{E}[Z_nY_{2,n}] + \mathbb{E}[Y_{2,n}]$. The lemma then follows by combining the above bounds for (2.9) and (2.10), and observing that $\lim_{n\to\infty}\frac{2e^{2T^{-1}}T^{-(1-x)}(\frac{n}{\lambda_n})^{\frac{1-x}{2}}}{1-x} = \frac{2e^{2T^{-1}}T^{-(1-x)}}{1-x}$, $\lim_{n\to\infty}\lambda_n^{\frac{x-1}{2}}C_{T,x} = 0$, and $\lim_{n\to\infty}\lambda_n^{-\frac{1}{2}}\mathbb{E}[Z_nY_{2,n}] = 0$. □

We now explicitly compute the limiting values of all expressions appearing in Lemma 6, allowing us to compute $\lim_{n\to\infty}\big(z_n(x)-\lambda_n\big)\lambda_n^{-\frac{1}{2}}$.

Lemma 7.

$$\lim_{n\to\infty}\big(z_n(x)-\lambda_n\big)\lambda_n^{-\frac{1}{2}} = \frac{\int_{-\infty}^{B}(B-z)^{1-x}e^{-\frac{z^2}{2}}dz}{\int_{-\infty}^{B}(B-z)^{-x}e^{-\frac{z^2}{2}}dz}. \quad (2.15)$$

Proof. It is easily verified that $\{Y_{1,n}\}, \{Y_{2,n}\}$, and $\{Z_nY_{2,n}\}$ are all sequences of uniformly integrable r.v.s. Let $f_1(y) \triangleq (B-y)^{1-x}I(y \le B), f_2(y) \triangleq (B-y)^{-x}I(y \le B-T^{-1})$, and $f_3(y) \triangleq y(B-y)^{-x}I(y \le B-T^{-1})$. Let N denote a normal r.v. with mean 0 and variance 1. It may be easily verified that $\{Y_{1,n}\}, \{Y_{2,n}\}$, and $\{Z_nY_{2,n}\}$ converge in distribution to $f_1(N), f_2(N), f_3(N)$ respectively, where all three convergences follow from the well-known convergence of the scaled Poisson r.v. Z_n to N. It thus follows from the uniform integrability of all three sequences that $\lim_{n\to\infty}\mathbb{E}[Y_{1,n}] = \mathbb{E}[f_1(N)] = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{B}(B-z)^{1-x}e^{-\frac{z^2}{2}}dz$, $\lim_{n\to\infty}\mathbb{E}[Y_{2,n}] = \mathbb{E}[f_2(N)] = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{B-T^{-1}}(B-$

$z)^{-x}e^{-\frac{z^2}{2}}dz$, and $\lim_{n\to\infty}\mathbb{E}[Z_nY_{2,n}] = \mathbb{E}[f_3(N)] = (2\pi)^{-\frac{1}{2}}\int_{-\infty}^{B-T^{-1}} z(B-z)^{-x}e^{-\frac{z^2}{2}}dz$. The lemma follows from plugging the above limits into Lemma 6, and letting $T \to \infty$. □

We now complete the proof of Proposition 2 by relating the integrals appearing in (2.15) to the parabolic cylinder functions. First, it will be useful to state some additional properties of the parabolic cylinder functions. $D_x(z)$ is an entire function of z for fixed x ([38], Chapter 8, Section 8.2) and an entire function of x for fixed z [24]. It is stated in [50] that for all $x, z \in \mathbb{R}$,

$$D_{x+1}(z) - zD_x(z) + xD_{x-1}(z) = 0; \tag{2.16}$$

$$\frac{d}{dz}D_x(z) + \frac{1}{2}zD_x(z) - xD_{x-1}(z) = 0. \tag{2.17}$$

With these properties in hand, we now complete the proof of Proposition 2.

Proof of Proposition 2. Note that the r.h.s. of (2.15) equals

$$\frac{\int_0^\infty t^{1-x}e^{-\frac{(B-t)^2}{2}}dt}{\int_0^\infty t^{-x}e^{-\frac{(B-t)^2}{2}}dt} = \frac{\Gamma(2-x)\frac{\left(D_x(-B)+BD_{x-1}(-B)\right)}{1-x}}{\Gamma(1-x)D_{x-1}(-B)} = \upsilon(x,-B)+B$$

by (2.1) and (2.16), since $\frac{\Gamma(2-x)}{\Gamma(1-x)} = (1-x)$, completing the proof. □

We now complete the proofs of Theorem 6 and Corollary 2.

Proof of Theorem 6. Since by definition $\Psi_n(x) = z_n(x) - a_n(x)$, Theorem 6 follows immediately from Lemma 4 and Proposition 2. □

Proof of Corollary 2. Suppose $0 < \epsilon < \min(\frac{B^2}{4}, 1-\frac{B^2}{4})$. Then for all sufficiently large n, $0 < \frac{B^2}{4} - \epsilon < (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < \min(1-\epsilon, \frac{B^2}{4}+\epsilon)$ by Lemma 3, and $\frac{B^2}{4} < 1-\epsilon$. By Lemma 1.ii and (2.5), we have that $\big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$ is a continuous, strictly decreasing function of x on $(-\infty, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 + \epsilon]$. It follows that for all sufficiently large n, $\big(z_n(\frac{B^2}{4} + \epsilon) - \lambda_n\big)\lambda_n^{-\frac{1}{2}} \leq \left(z_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) - \lambda_n\right)\lambda_n^{-\frac{1}{2}} \leq \big(z_n(\frac{B^2}{4} - \epsilon) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$. Thus by Proposition 2, for all sufficiently small $\epsilon > 0$,

$$\begin{aligned} z_\infty(\frac{B^2}{4} + \epsilon) &\leq \liminf_{n\to\infty} \left(z_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) - \lambda_n\right)\lambda_n^{-\frac{1}{2}} \\ &\leq \limsup_{n\to\infty} \left(z_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) - \lambda_n\right)\lambda_n^{-\frac{1}{2}} \\ &\leq z_\infty(\frac{B^2}{4} - \epsilon). \end{aligned} \tag{2.18}$$

To proceed, we will now prove that $z_\infty(x)$ has certain continuity properties. In particular,

$$z_\infty(x) \text{ is a differentiable, non-increasing function of } x \text{ on } (-\infty, 1]. \tag{2.19}$$

Indeed, by (2.3), $D_{x-1}(-B) > 0$ for $x \leq 1$. The differentiability of $z_\infty(x)$ on $(-\infty, 1]$ then follows from the fact that $D_x(-B)$ is an entire function of x. That $z_\infty(x)$ is non-increasing follows from Lemma 1.ii, (2.5), and Proposition 2.

The continuity of $z_\infty(x)$ in a neighborhood of $\frac{B^2}{4}$ (guaranteed by (2.19) since $0 < \frac{B^2}{4} < 1$), along with (2.18), thus implies that $\lim_{n\to\infty}\left(z_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) - \lambda_n\right)\lambda_n^{-\frac{1}{2}} = z_\infty(\frac{B^2}{4})$. By (2.7), $\left(a_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) - \lambda_n\right)\lambda_n^{-\frac{1}{2}} = n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}}$. Thus by Lemma 3, $\lim_{n\to\infty}\left(a_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) - \lambda_n\right)\lambda_n^{-\frac{1}{2}} = \frac{B}{2}$, and the corollary follows since $\zeta(B) = z_\infty(\frac{B^2}{4}) - \frac{B}{2}$. □

2.5 Asymptotic Analysis of the Function $Z^+(\Psi_n)$

In this section we derive the asymptotics of $Z^+(\Psi_n)$. In particular, we prove

Theorem 7. *If $B < 2$ and $\zeta(B) < 0$, or $B \geq 2$, then* $\lim_{n\to\infty} Z^+(\Psi_n) = Z^+(\Psi_\infty)$.

We begin by establishing some additional properties of $\Psi_\infty(x)$.

Lemma 8. *$\Psi_\infty(x)$ is a continuous function of x on $\big(-\infty, \min(1, \frac{B^2}{4})\big)$, left-continuous at $\min(1, \frac{B^2}{4})$, and a differentiable function of x on $\big(-\infty, \min(1, \frac{B^2}{4})\big)$, satisfying $\frac{d}{dx}\Psi_\infty(x) \leq -(B^2 - 4x)^{-\frac{1}{2}}$ for $x < \min(1, \frac{B^2}{4})$.*

Proof. That a_∞ is differentiable on $(-\infty, \frac{B^2}{4})$, left-continuous at $\frac{B^2}{4}$, and satisfies $\frac{d}{dx}a_\infty(x) = (B^2 - 4x)^{-\frac{1}{2}}$ for $x < \frac{B^2}{4}$, follows from elementary calculus. The lemma then follows from (2.19) and the fact that $\Psi_\infty(x) = z_\infty(x) - a_\infty(x)$. □

We now prove some additional properties of $Z^+(\Psi_\infty)$.

Lemma 9. *If $B < 2$ and $\zeta(B) < 0$, or $B \geq 2$, then: $\Psi_\infty(x)$ has a unique zero $Z^+(\Psi_\infty) \in \big(0, \min(1, \frac{B^2}{4})\big)$; $\Psi_\infty(x) > 0$ on $[0, Z^+(\Psi_\infty))$; and*
$\Psi_\infty(x) < 0$ on $\big(Z^+(\Psi_\infty), \min(1, \frac{B^2}{4})\big]$. Alternatively, if $B < 2$ and $\zeta(B) = 0$, then: $\Psi_\infty(x) > 0$ on $\big[0, \min(1, \frac{B^2}{4})\big)$, and $Z^+(\Psi_\infty) = \frac{B^2}{4}$.

Proof. We first treat the case $B < 2$ and $\zeta(B) < 0$, or $B \geq 2$. By Lemma 8, $\Psi_\infty(x)$ is a strictly decreasing function of x on $\big[0, \min(1, \frac{B^2}{4})\big)$, continuous on $\big[0, \min(1, \frac{B^2}{4})\big)$, and left-continuous at $\min(1, \frac{B^2}{4})$. $\Psi_\infty(0) > 0$, since $\Psi_\infty(0) = v(0, -B) + B$, and by (2.2), $v(0, -B) > 0$.

Also, $\Psi_\infty(\min(1, \frac{B^2}{4})) < 0$. Indeed, if $B \geq 2$, $\min(1, \frac{B^2}{4}) = 1$. Furthermore, $\Psi_\infty(1) < 0$, since by (2.2), $\Psi_\infty(1) = -B + \frac{1}{2}\big(B + (B^2 - 4)^{\frac{1}{2}}\big) < 0$. If $B < 2$, then $\min(1, \frac{B^2}{4}) = \frac{B^2}{4}$, and $\Psi_\infty(\frac{B^2}{4}) = \zeta(B) < 0$. Combining the cases $B \leq 2$ and $B > 2$ demonstrates that $\Psi_\infty(\min(1, \frac{B^2}{4})) < 0$, and combining the above completes the proof for the case $B < 2$ and $\zeta(B) < 0$, or $B \geq 2$.

We now treat the case $B < 2$ and $\zeta(B) = 0$. In this case, the lemma follows from the fact that $\Psi_\infty(x)$ is strictly decreasing on $\big(-\infty, \min(1, \frac{B^2}{4})\big)$ and left-continuous at $\frac{B^2}{4}$ by Lemma 8, and $\zeta(B) = 0$. □

Proof of Theorem 7. We first show that for all sufficiently large n, $Z^+(\Psi_n)$ is the unique zero of Ψ_n in the interval $\big(0, \min(1, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)$. Indeed, if $B < 2$ and $\zeta(B) < 0$, then for all sufficiently large n, $\min\big(1, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2)\big) = (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$ by Lemma 3 and $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$ by Corollary 2. It follows from Lemma 2.i that for all sufficiently large n, $Z^+(\Psi_n)$ is the unique zero of Ψ_n in the interval $\min\big(1, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)$. If instead $B \geq 2$, then for all sufficiently large n, $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 \geq 1$ and $\min\big(1, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2)\big) = 1$ by Lemma 3. It follows from Lemma 2.iii that for all sufficiently large n, $Z^+(\Psi_n)$ is the unique zero of Ψ_n in the interval $(0, \min\big(1, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2)\big)$.

Now suppose for contradiction that $\liminf_{n\to\infty} Z^+(\Psi_n) < Z^+(\Psi_\infty)$. By Lemma 9, $Z^+(\Psi_\infty) \in \big(0, \min(1, \frac{B^2}{4})\big)$, and $\Psi_\infty(x) > 0$ on $\big[0, Z^+(\Psi_\infty)\big)$. It follows that there exists $\epsilon > 0$ such that $0 < \liminf_{n\to\infty} Z^+(\Psi_n) + \epsilon < \min(1, \frac{B^2}{4})$, and $\Psi_\infty\big(\liminf_{n\to\infty} Z^+(\Psi_n) + \epsilon\big) > 0$. Thus by Theorem 6, for all sufficiently large n, $\Psi_n(\liminf_{n\to\infty} Z^+(\Psi_n) + \epsilon) > 0$. By the monotonicity of $\Psi_n(x)$ (from Lemma 1.iv), it follows that for all sufficiently large n, $\Psi_n(x) > 0$ on $(-\infty, \liminf_{n\to\infty} Z^+(\Psi_n) + \epsilon)$. But by the definition of lim inf, there exists an infinite strictly increasing sequence of integers $\{n_i\}$ s.t. $Z^+(\Psi_{n_i}) < \liminf_{n\to\infty} Z^+(\Psi_n) + \epsilon$ for all i. Thus for all sufficiently large i, $\Psi_{n_i}\big(Z^+(\Psi_{n_i})\big) > 0$. This is a contradiction, since from definitions $\Psi_{n_i}\big(Z^+(\Psi_{n_i})\big) = 0$ for all i.

Alternatively, suppose for contradiction that $\limsup_{n\to\infty} Z^+(\Psi_n) > Z^+(\Psi_\infty)$. By Lemma 9, $Z^+(\Psi_\infty) \in \big(0, \min(1, \frac{B^2}{4})\big)$, and $\Psi_\infty(x) < 0$ on $\big(Z^+(\Psi_\infty), \min(1, \frac{B^2}{4})\big]$. It follows that there exists $\epsilon > 0$ such that $0 < \limsup_{n\to\infty} Z^+(\Psi_n) - \epsilon < \min(1, \frac{B^2}{4})$, and $\Psi_\infty\big(\limsup_{n\to\infty} Z^+(\Psi_n) - \epsilon\big) < 0$. Thus by Theorem 6, for all sufficiently large n, $\Psi_n\big(\limsup_{n\to\infty} Z^+(\Psi_n) - \epsilon\big) < 0$. By the monotonicity of $\Psi_n(x)$ (from Lemma 1.iv), it follows that for all sufficiently large n, $\Psi_n(x) < 0$ on $\Big(\limsup_{n\to\infty} Z^+(\Psi_n) - \epsilon, \min\big(1, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)\Big)$.

By the definition of lim sup, there exists an infinite strictly increasing sequence of integers $\{n_i\}$ s.t. $Z^+(\Psi_{n_i}) > \limsup_{n\to\infty} Z^+(\Psi_n) - \epsilon$ for all i. Combining the above, we find that for all sufficiently large i, $\Psi_{n_i}\big(Z^+(\Psi_{n_i})\big) < 0$. But this is a contradiction, since from definitions $\Psi_{n_i}\big(Z^+(\Psi_{n_i})\big) = 0$ for all i. We conclude that for the case $B < 2$ and $\zeta(B) < 0$, or $B \geq 2$, $\liminf_{n\to\infty} Z^+(\Psi_n) = \limsup_{n\to\infty} Z^+(\Psi_n) = Z^+(\Psi_\infty)$. □

2.6 The Zeros of $\zeta(B)$

In this section we characterize the set of $B \in (0,2)$ for which $\zeta(B) \leq 0$, which will allow us to use Theorem 7 effectively. In particular, we prove

Theorem 8. *$B^* \in [2^{\frac{1}{2}}, 2)$. $\zeta(B) > 0$ on $[0, B^*)$, and $\zeta(B) < 0$ on $(B^*, 2]$.*

We also complete the proof of Proposition 1. We begin by proving some properties of the derivatives of $\upsilon(x,y)$, which will then enable us to analyze $\zeta'(B) \triangleq \frac{d}{dB}\zeta(B)$ using the multivariate chain rule.

Lemma 10. *$D_{x-1}(y) \in (0,\infty)$ for $x \leq 1$ and all y. Also, $\upsilon(x,y)$ is a smooth function of y for $x \leq 1$, and a smooth function of x for $x \leq 1$ and all y. Furthermore, $\frac{d}{dy}\upsilon(x,y) = x + \upsilon^2(x,y) - y\upsilon(x,y)$.*

Proof. Note that if $x \leq 1$ then $D_{x-1}(y) > 0$ for all y by (2.3), from which the first part of the lemma follows. Since $D_x(y)$ is an entire function of y for each fixed x, it follows that for $x \leq 1$, $\upsilon(x,y)$ is a smooth function of y. The second part of the lemma then follows from the fact that $D_x(y)$ is an entire function of x for all y, and (2.3). The final part of the lemma then follows from the chain rule, (2.17), and (2.16). □

We now prove some properties, including existence, of $\varphi'(B) \triangleq \frac{d}{dB}\varphi(B)$, by applying the chain rule to Lemma 10.

Lemma 11. *$D_{\frac{B^2}{4}-1}(-B) \in (0,\infty)$ on $(-\infty, 2]$. Furthermore, $\varphi(B)$ is a differentiable function of B on $(-\infty, 2]$, and on $(-\infty, 2]$ we have*

$$\varphi'(B) = \frac{B}{2}\frac{dv}{dx}(\frac{B^2}{4}, -B) - \frac{B^2}{4} - \varphi^2(B) - B\varphi(B).$$

Proof. $D_{\frac{B^2}{4}-1}(-B) > 0$ for $B \leq 2$ by (2.3). The lemma then follows from the multivariate chain rule and Lemma 10. □

Although it seems clear from Figure 2-1 that $\zeta(B)$ is strictly negative on $(B^*, 2)$, the formal proof of this fact is somewhat involved, and will necessitate careful arguing about $\varphi'(B)$. Due to the fact that $\frac{d}{dx}D_x(y)$ has no simple analytic form (as opposed to $\frac{d}{dy}D_x(y)$, see (2.17)), to proceed we will have to derive good bounds for $\frac{d}{dx}v(x,y)$. Our bounds will rely on the concavity of certain functions. We begin by proving that

Lemma 12. *For any fixed $y < 0$, $v(x,y)$ is a concave function of x on $(0,1)$.*

Proof. We begin by demonstrating that $z_{n,k}(x)$ is a twice-differentiable concave function of x on $(0,1)$ for all $k \leq n$, which will imply that $z_\infty(x)$, and ultimately $v(x,y)$ are concave by taking limits. We proceed by induction on k. For the base case, consider $k = 1$. By definitions, $z_{n,1}(x) = \lambda_n + 1 - x$, and the assertion is trivial. Now, let us assume the statement is true for $j = 1, \ldots, k-1$ with $k - 1 \leq n - 1$. By definitions, $f_{n,k}(x) = (\lambda_n + k - x)f_{n,k-1}(x) - \lambda_n(k-1)f_{n,k-2}(x)$. Since by (2.5), $Z(f_{n,n-1}) > 1$, we have by Lemma 1.i that $f_{n,k-1}(x) > 0$ and $f_{n,k-2}(x) > 0$, from which it follows that $z_{n,k-1}(x) > 0$ and $z_{n,k}(x) = (\lambda_n + k - x) - \lambda_n(k-1)(z_{n,k-1}(x))^{-1}$. Thus

$$\frac{d^2}{dx^2}z_{n,k}(x) = \lambda_n(k-1)\Big(- 2z_{n,k-1}(x)^{-3}\big(\frac{d}{dx}z_{n,k-1}(x)\big)^2 + z_{n,k-1}(x)^{-2}\frac{d^2}{dx^2}z_{n,k-1}(x)\Big). \tag{2.20}$$

Since $z_{n,k-1}(x) > 0$, $\big(\frac{d}{dx}z_{n,k-1}(x)\big)^2 > 0$, and by the induction hypothesis $\frac{d^2}{dx^2}z_{n,k-1}(x) \leq 0$, it follows from (2.20) that $z_{n,k}(x)$ is twice-differentiable on $(0,1)$ and satisfies

$\frac{d^2}{dx^2}z_{n,k}(x) \leq 0$ (concavity), proving the induction.

It follows that for any fixed $B > 0$, $z_\infty(x)$ is a concave function of x on $(0,1)$, due to Proposition 2, and the fact that pointwise limits of concave functions are concave. The lemma follows, since $v(x,y) = z_\infty(x) - B$. □

We now use the concavity of $v(x,y)$ and the associated monotonicity of $\frac{d}{dx}v(x,y)$ to prove that

Lemma 13. *For any fixed $y < 0$ and $x_0 \in (0,1)$, $\frac{dv}{dx}(x_0,y) < x_0^{-1}v(x_0,y)$.*

Proof. By Lemma 10 and the Mean Value Theorem, there exists $c \in (0,x_0)$ such that $\frac{dv}{dx}(c,y) = x_0^{-1}\big(v(x_0,y) - v(0,y)\big)$. The lemma then follows from concavity, since $v(0,y) > 0$ by (2.2). □

We now plug our bounds on $\frac{d}{dx}v(x,y)$ from Lemma 13 into Lemma 11 to prove that

Lemma 14. *For $0 < B < 2$,*

$$\varphi'(B) < (\frac{2}{B} - B)\varphi(B) - \frac{B^2}{4} - \varphi^2(B) \leq \frac{1}{B^2} - 1.$$

Proof. By Lemma 11 and Lemma 13, for $0 < B < 2$,

$$\begin{aligned}\varphi'(B) &< \frac{B}{2}(\frac{B^2}{4})^{-1}\varphi(B) - \frac{B^2}{4} - \varphi^2(B) - B\varphi(B)\\ &= (\frac{2}{B} - B)\varphi(B) - \frac{B^2}{4} - \varphi^2(B).\end{aligned}$$

This proves the first inequality. Furthermore, it follows that there exists $x_B \in \mathbb{R}$ (we may take $x_B = \varphi(B)$) such that $\varphi'(B) \leq \alpha(B, x_B) \triangleq (\frac{2}{B} - B)x_B - x_B^2 - \frac{B^2}{4}$. Since by elementary calculus, for any fixed $B > 0$, the function $\alpha(B,x)$ attains its maximum at $x = \frac{(\frac{2}{B}-B)}{2}$, we find that for $B \in (0,2)$,

$$\varphi'(B) \leq (\frac{2}{B} - B)(\frac{(\frac{2}{B} - B)}{2}) - (\frac{(\frac{2}{B} - B)}{2})^2 - \frac{B^2}{4} = \frac{1}{B^2} - 1.$$

□

We are now in a position to study the zeros of $\zeta(B)$.

Lemma 15. *$\zeta(B) > 0$ on $[0, B^*)$, and $B^* \in (0,2)$.*

Proof. By (2.2), $\zeta(0) = \left(2^{\frac{1}{2}} \int_0^\infty e^{-t^2} dt\right)^{-1} > 0$, and $\zeta(2) = \frac{-2e^{-1}}{e^{-1}} + 1 = -1 < 0$. By Lemma 11 and definitions, $\zeta(B)$ is a differentiable (and continuous) function of B on $(-\infty, 2]$. It follows from the Intermediate Value Theorem that $B^* \in (0,2)$, and $\zeta(B) > 0$ on $[0, B^*)$. □

Lemma 16. *$B^* \in [2^{\frac{1}{2}}, 2)$.*

Proof. It is proven in [104] Theorem 4.1 i. that

$$\gamma_n \geq \inf_{k \geq 1} \Big(\lambda_n + \min(k,n) - \lambda_n^{\frac{1}{2}}\big(\overset{\frac{1}{2}}{\min}(k-1,n) + \overset{\frac{1}{2}}{\min}(k,n)\ \big)\Big). \tag{2.21}$$

Note that for $1 \leq k \leq n$, $\lambda_n + \min(k,n) - \lambda_n^{\frac{1}{2}}\big(\ \min^{\frac{1}{2}}(k-1,n) + \min^{\frac{1}{2}}(k,n)\ \big)$ equals

$$(\lambda_n^{\frac{1}{2}} - k^{\frac{1}{2}})^2 + \frac{\lambda_n^{\frac{1}{2}}}{k^{\frac{1}{2}} + (k-1)^{\frac{1}{2}}} \geq \frac{1}{2}\frac{\lambda_n^{\frac{1}{2}}}{n^{\frac{1}{2}}}. \tag{2.22}$$

For all $k \geq n+1$, the r.h.s. of (2.21) equals $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. Combining this fact with (2.21) and (2.22), we find that

$$\gamma_n \geq \min\big(\frac{1}{2}\frac{\lambda_n^{\frac{1}{2}}}{n^{\frac{1}{2}}}, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big). \tag{2.23}$$

By Lemma 3, $\lim_{n\to\infty}(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 = \frac{B^2}{4}$. Trivially, $\lim_{n\to\infty} \frac{1}{2}\frac{\lambda_n^{\frac{1}{2}}}{n^{\frac{1}{2}}} = \frac{1}{2}$. Thus for any $B < 2^{\frac{1}{2}}$ and all sufficiently large n, $\frac{1}{2}\frac{\lambda_n^{\frac{1}{2}}}{n^{\frac{1}{2}}} > (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. It follows that $\gamma_n \geq (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$ for all sufficiently large n by (2.23).

Now, suppose for contradiction that $B^* < 2^{\frac{1}{2}}$. By Lemma 14 and the fact that $B^* \in (0,2)$ by Lemma 15, we have that $\varphi(B)$ is differentiable at B^*, and $\varphi'(B^*)$ is strictly less than

$$(\frac{2}{B^*} - B^*)(-\frac{B^*}{2}) - \frac{B^{*2}}{4} - (-\frac{B^*}{2})^2 \; = \; -1,$$

since $\varphi(B^*) = -\frac{B^*}{2}$. It follows that $\zeta'(B^*) < 0$, since $\zeta'(B^*) = \varphi'(B^*) + \frac{1}{2}$. Thus there exists $\epsilon > 0$ s.t. $\zeta(B) < 0$ on $(B^*, B^* + \epsilon)$ since $\zeta(B^*) = 0$. It follows that there exists $B' \in (0, 2^{\frac{1}{2}})$ such that $\zeta(B') < 0$. Thus if we define all relevant functions in terms of B', $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$ for all sufficiently large n by Corollary 2. Also, $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 1$ for all sufficiently large n by Lemma 3. Thus by Lemma 2.i, for all sufficiently large n, $\gamma_n < (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. But this is a contradiction, since we have already shown that $B' < 2^{\frac{1}{2}}$ implies that $\gamma_n \geq (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$ for all sufficiently large n, completing the proof. □

Proof of Theorem 8. In light of Lemma 15 and Lemma 16, all that remains to be shown is that $\zeta(B) < 0$ on $(B^*, 2)$. In light of Lemma 16, it would suffice to prove that $\zeta(B)$ is strictly decreasing on $(2^{\frac{1}{2}}, 2)$. But for $B \in (2^{\frac{1}{2}}, 2)$, $\zeta'(B)$ equals

$$\varphi'(B) + \frac{1}{2} \; < \; \frac{1}{B^2} - 1 + \frac{1}{2} \; = 0$$

by Lemma 14. This demonstrates that $\zeta(B)$ is strictly decreasing on $(2^{\frac{1}{2}}, 2)$, concluding the proof of Theorem 8. □

Proof of Proposition 1. That $B^* \in [2^{\frac{1}{2}}, 2)$ follows immediately from Theorem 8. That $Z^+(\Psi_\infty) < \min(1, \frac{B^2}{4})$ for $B > B^*$ follows immediately from Theorem 8 and Lemma 9. □

2.7 Proof of Main Results

In this section we complete the proofs of our main results, Theorem 4 and Corollary 1.

Proof of Theorem 4. First, suppose $0 < B < B^*$. Then $B < 2$ by Theorem 8, and thus $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 1$ for all sufficiently large n by Lemma 3. $\zeta(B) > 0$ by Theorem 8, and thus since $B < 2$, $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) > 0$ for all sufficiently large n by Corollary 2. It follows from Lemma 2.ii that $\gamma_n = (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$ for all sufficiently large n. That $\lim_{n\to\infty} \gamma_n = \frac{B^2}{4}$ then follows from Lemma 3.

Now, suppose $B = B^*$. Since $B^* < 2$ by Theorem 8, it follows from Lemma 3 that $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 1$ for all sufficiently large n. By Lemma 2, for all sufficiently large n, either $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$, in which case $\gamma_n = Z^+(\Psi_n)$ is the unique zero of Ψ_n in the interval $\big(0, (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)$, or $\gamma_n = (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. Let $\{n_i\}$ denote the subsequence of $\{n\}$ for which $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$. If $\{n_i\}$ is a finite set, then trivially $\gamma_n = (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$ for all sufficiently large n. That $\lim_{n\to\infty} \gamma_n = \frac{B^{*2}}{4}$ then follows from Lemma 3.

Alternatively, suppose $\{n_i\}$ is an infinite set. Suppose for contradiction that $\liminf_{i\to\infty} Z^+(\Psi_{n_i}) < \frac{B^{*2}}{4}$. Note that by Lemma 9, $Z^+(\Psi_\infty) = \frac{B^{*2}}{4}$, and $\Psi_\infty(x) > 0$ on $[0, \frac{B^{*2}}{4})$. It follows that there exists $\epsilon > 0$ such that $0 < \liminf_{i\to\infty} Z^+(\Psi_{n_i}) + \epsilon < \frac{B^{*2}}{4} < 1$ and $\Psi_\infty\big(\liminf_{i\to\infty} Z^+(\Psi_{n_i}) + \epsilon\big) > 0$. It follows from Lemma 3 and Theorem 6 that for all sufficiently large i, $0 < \liminf_{i\to\infty} Z^+(\Psi_{n_i}) + \epsilon < (n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2 < 1$ and $\Psi_{n_i}\big(\liminf_{i\to\infty} Z^+(\Psi_{n_i}) + \epsilon\big) > 0$. By the monotonicity of $\Psi_{n_i}(x)$ (from Lemma 1.iv), it follows that for all sufficiently large i, $\Psi_{n_i}(x) > 0$ on $\big(-\infty, \liminf_{i\to\infty} Z^+(\Psi_{n_i}) + \epsilon\big)$. But by the definition of $\liminf$, there exists an infinite strictly increasing sequence of integers $\{n_i'\}$ s.t. $Z^+(\Psi_{n_i'}) < \liminf_{i\to\infty} Z^+(\Psi_{n_i}) + \epsilon$ for all i. Thus for all sufficiently large i, $\Psi_{n_i'}\big(Z^+(\Psi_{n_i'})\big) > 0$. This is a contradiction, since $\Psi_{n_i'}\big(Z^+(\Psi_{n_i'})\big) = 0$ for all i. This proves that $\liminf_{i\to\infty} Z^+(\Psi_{n_i}) \geq \frac{B^{*2}}{4}$.

Note that $(n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2 < 1$ for all sufficiently large i by Lemma 3 and Proposition

1. It follows from Lemma 2.i that for all sufficiently large i, $Z^+(\Psi_{n_i}) < (n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2$, and thus $\limsup_{i\to\infty} Z^+(\Psi_{n_i}) \leq \frac{B^{*2}}{4}$ by Lemma 3. Combining the above, we find that $\lim_{i\to\infty} Z^+(\Psi_{n_i}) = \frac{B^{*2}}{4}$. Thus γ_n alternates between two sequences, that of the $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$ and that of the $Z^+(\Psi_{n_i})$, both of which converge to $\frac{B^{*2}}{4}$, proving that $\lim_{n\to\infty} \gamma_n = \frac{B^{*2}}{4} = Z^+(\Psi_\infty)$.

Now, suppose $B \in (B^*, 2)$. Then $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 1$ for all sufficiently large n by Lemma 3. $\zeta(B) < 0$ by Theorem 8, and thus since $B < 2$, $\Psi_n\big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big) < 0$ for all sufficiently large n by Corollary 2. It follows from Lemma 2.i that $\gamma_n = Z^+(\Psi_n)$ for all sufficiently large n. That $\lim_{n\to\infty} \gamma_n = Z^+(\Psi_\infty)$ then follows from Theorem 7.

Now, suppose $B \geq 2$. Then $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 > 1$ for all sufficiently large n by Lemma 3. It follows from Lemma 2.iii that $\gamma_n = Z^+(\Psi_n)$ for all sufficiently large n. That $\lim_{n\to\infty} \gamma_n = Z^+(\Psi_\infty)$ then follows from Theorem 7. This treats all cases, and we conclude that $\lim_{n\to\infty} \gamma_n \overset{\Delta}{=} \gamma_B$ exists for all $B > 0$. We also conclude that for $0 < B \leq B^*$, $\gamma_B = \frac{B^2}{4}$; for $B \geq B^*$, $\gamma_B = Z^+(\Psi_\infty)$. □

Proof of Corollary 1. Suppose for contradiction that $\liminf_{n\to\infty} n^{\frac{1}{2}}(1 - \rho_n^*) < B^*$. Then there exists $\epsilon > 0$ and an infinite strictly increasing sequence of integers $\{n_i\}$ s.t. $n_i^{\frac{1}{2}}(1 - \rho^*_{n_i}) < B^* - \epsilon$, and thus $\rho^*_{n_i} > 1 - (B^* - \epsilon)n_i^{-\frac{1}{2}}$. Consider the sequence $\{Z_i\}$ of continuous time markov chains, in which Z_i is an $M/M/n_i$ queueing system with $\lambda_{n_i} = n_i - (B^* - \epsilon)n_i^{\frac{1}{2}}$, $\mu = 1$. Since $B^* - \epsilon < B^*$, it follows from Theorem 8 that $\zeta(B^* - \epsilon) > 0$ and $B^* - \epsilon < 2$. Now, let us define all relevant functions (e.g. $\Psi_{n_i}(x)$) w.r.t. $B^* - \epsilon$. It then follows from Corollary 2 and Lemma 3 that for all sufficiently large i, $\Psi_{n_i}\big((n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2\big) > 0$, and $(n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2 < 1$. Thus by Lemma 2.ii, $\gamma_{n_i} = (n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2$ for all sufficiently large i. However, note that by assumption $\frac{\lambda_{n_i}}{n_i\mu} = 1 - (B^* - \epsilon)n_i^{-\frac{1}{2}} < \rho^*_{n_i}$ for all i. But this is a contradiction, since by Theorem 3, $\frac{\lambda_{n_i}}{n_i\mu} < \rho^*_{n_i}$ implies that the spectral gap γ_{n_i} of Z_i is strictly less than $(n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2$.

Alternatively, suppose for contradiction that $\limsup_{n\to\infty} n^{\frac{1}{2}}(1 - \rho_n^*) > B^*$. Then there exists $\epsilon \in (0, 2-B^*)$ and an infinite strictly increasing sequence of integers $\{n_i\}$ s.t. $n_i^{\frac{1}{2}}(1-\rho^*_{n_i}) > B^*+\epsilon$, and thus $\rho^*_{n_i} < 1-(B^*+\epsilon)n_i^{-\frac{1}{2}}$. Consider the sequence $\{Z_i\}$ of continuous time Markov chains, in which Z_i is an $M/M/n_i$ queueing system with $\lambda_{n_i} = n_i - (B^*+\epsilon)n_i^{\frac{1}{2}}$, $\mu = 1$. Since $B^*+\epsilon \in (B^*, 2)$, it follows from Theorem 8 that $\zeta(B^*+\epsilon) < 0$. Now, let us define all relevant functions w.r.t. $B^*+\epsilon$. It then follows from Corollary 2 and Lemma 3 that for all sufficiently large i, $\Psi_{n_i}\big((n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2\big) < 0$, and $(n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2 < 1$. Thus by Lemma 2.i, $\gamma_{n_i} = Z^+(\Psi_{n_i}) < (n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2$ for all sufficiently large i. However, note that by assumption $\frac{\lambda_{n_i}}{n_i\mu} = 1-(B^*+\epsilon)n_i^{-\frac{1}{2}} > \rho^*_{n_i}$ for all i. But this is a contradiction, since by Theorem 3, $\frac{\lambda_{n_i}}{n_i\mu} > \rho^*_{n_i}$ implies that the spectral gap γ_{n_i} of Z_i is equal to $(n_i^{\frac{1}{2}} - \lambda_{n_i}^{\frac{1}{2}})^2$. It follows that $\liminf_{n\to\infty} n^{\frac{1}{2}}(1-\rho^*_n) = \limsup_{n\to\infty} n^{\frac{1}{2}}(1-\rho^*_n) = B^*$, completing the proof. □

2.8 Conclusion and Open Questions

In this chapter we studied the rate of convergence to stationarity of the $M/M/n$ queue in the H-W Regime. We explicitly computed the limit of the exponential rate of convergence to stationarity, i.e. the spectral gap. We proved that there is an interesting phase transition in the system's behavior, occuring when the excess parameter B reaches $B^* \approx 1.85772$. For $B < B^*$, the exponential rate of convergence is $\frac{B^2}{4}$; above B^* it is the solution to an equation involving the parabolic cylinder functions. We showed that this transition asymptotically characterizes a phenomenon previously observed to occur for fixed n, unifying and simplifying several earlier lines of work.

This work leaves several interesting directions for future research. There are many open questions related to the interaction between weak convergence and convergence to stationarity. Although our results and those of [70] show that for the $M/M/n$ queue in the H-W regime there is an 'interchange of limits' in this regard, namely the limiting rate of convergence equals the rate of convergence of the limit, it is unknown to what extent such an interchange must hold in general. It would also be interesting to prove that a phase transition occurs in other related models, and for preliminary results along these lines the reader is referred to the recent paper [71].

2.9 Appendix

2.9.1 Proof of Lemma 5

In this subsection, we complete the proof of Lemma 5.

Proof of Lemma 5. By (2.5), $f_{n,n-1}(x) > 0$ on $(0,1)$, and thus by (2.4), $\big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$ equals

$$\begin{aligned}&(\frac{\sum_{k=0}^{n}\binom{n}{k}\lambda_n^k\prod_{j=1}^{n-k}(j-x)}{\sum_{k=0}^{n-1}\binom{n-1}{k}\lambda_n^k\prod_{j=1}^{n-1-k}(j-x)} - \lambda_n)\lambda_n^{-\frac{1}{2}}\\ &= \frac{\sum_{k=0}^{n}\binom{n}{k}\lambda_n^k\prod_{j=1}^{n-k}(j-x) - \lambda_n\sum_{k=0}^{n-1}\binom{n-1}{k}\lambda_n^k\prod_{j=1}^{n-1-k}(j-x)}{\lambda_n^{\frac{1}{2}}\sum_{k=0}^{n-1}\binom{n-1}{k}\lambda_n^k\prod_{j=1}^{n-1-k}(j-x)}. \qquad (2.24)\end{aligned}$$

Note that the numerator of (2.24) equals

$$\begin{aligned}&\prod_{j=1}^{n}(j-x) + \sum_{k=1}^{n}\binom{n}{k}\lambda_n^k\prod_{j=1}^{n-k}(j-x) - \sum_{k=0}^{n-1}\binom{n-1}{(k+1)-1}\lambda_n^{k+1}\prod_{j=1}^{n-(k+1)}(j-x)\\ &= (n-1)!\sum_{k=0}^{n}(n-k)\prod_{j=1}^{n-k}(1-\frac{x}{j})\,\frac{\lambda_n^k}{k!}; \qquad (2.25)\end{aligned}$$

and the denominator of (2.24) equals

$$\begin{aligned}\lambda_n^{-\frac{1}{2}}\sum_{k=0}^{n-1}\binom{n-1}{(k+1)-1}\lambda_n^{k+1}\prod_{j=1}^{n-(k+1)}(j-x) &= \lambda_n^{-\frac{1}{2}}\sum_{k=1}^{n}\frac{k}{n}\binom{n}{k}\lambda_n^{k}\prod_{j=1}^{n-k}(j-x)\\ &= \lambda_n^{-\frac{1}{2}}(n-1)!\sum_{k=0}^{n}k\prod_{j=1}^{n-k}(1-\frac{x}{j})\ \frac{\lambda_n^k}{k!} \quad (2.26)\end{aligned}$$

Combining (2.25) and (2.26), and normalizing both by $\frac{e^{-\lambda_n}}{(n-1)!}$, we find that $\big(z_n(x)-\lambda_n\big)\lambda_n^{-\frac{1}{2}}$ equals

$$\lambda_n^{\frac{1}{2}}\frac{\sum_{k=0}^{n}(n-k)\prod_{j=1}^{n-k}(1-\frac{x}{j})\ e^{-\lambda_n}\frac{\lambda_n^k}{k!}}{\sum_{k=0}^{n}k\prod_{j=1}^{n-k}(1-\frac{x}{j})\ e^{-\lambda_n}\frac{\lambda_n^k}{k!}}. \tag{2.27}$$

We now demonstrate that the numerator of (2.27) (ignoring the $\lambda_n^{\frac{1}{2}}$ prefactor) is bounded from below by

$$\prod_{j=1}^{T}(1-\frac{x}{j})\sum_{k=0}^{n-(T+1)}(n-k)\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\ e^{-\lambda_n}\frac{\lambda_n^k}{k!},$$

and bounded from above by

$$\prod_{j=1}^{T}(1-\frac{x}{j})(\sum_{k=0}^{n-(T+1)}(n-k)\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\ e^{-\lambda_n}\frac{\lambda_n^k}{k!}+\frac{(T+1)^2}{\prod_{j=1}^{T}(1-\frac{x}{j})n^{\frac{1}{2}}}).$$

Note that by a simple application of Stirling's Inequality, we find that

$$\lim_{n\to\infty}n^{\frac{1}{2}}e^{-\lambda_n}\frac{\lambda_n^{\lfloor\lambda_n\rfloor}}{\lfloor\lambda_n\rfloor!}=(2\pi^{\frac{1}{2}})^{-1}.$$

Thus we may w.l.o.g. assume that n is sufficiently large to ensure that $\frac{n}{\lambda_n}\leq 2$, $\lceil n-T^{-1}n^{\frac{1}{2}}\rceil+1\leq n-(T+1)$, $\lambda_n^{-\frac{1}{2}}\leq\frac{1}{2T}$, and $e^{-\lambda_n}\frac{\lambda_n^{\lfloor\lambda_n\rfloor}}{\lfloor\lambda_n\rfloor!}\leq n^{-\frac{1}{2}}$, and we assume this throughout.

The numerator of (2.27) equals

$$\prod_{j=1}^{T}(1-\frac{x}{j})\Big(\sum_{k=0}^{n-(T+1)}(n-k)\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\,e^{-\lambda_n}\frac{\lambda_n^k}{k!} \tag{2.28}$$

$$+(\prod_{j=1}^{T}(1-\frac{x}{j}))^{-1}\sum_{k=n-T}^{n}(n-k)\prod_{j=1}^{n-k}(1-\frac{x}{j})\,e^{-\lambda_n}\frac{\lambda_n^k}{k!}\Big).$$

The second summand in (2.28) is non-negative, yielding a trivial lower bound. We now obtain an upper bound on the same expression. Since the mode of a Poisson r.v. with mean λ_n is $\lfloor\lambda_n\rfloor$, and by assumption $e^{-\lambda_n}\frac{\lambda_n^{\lfloor\lambda_n\rfloor}}{\lfloor\lambda_n\rfloor!}\leq n^{-\frac{1}{2}}$, we have that for all $k\geq 0$,

$$e^{-\lambda_n}\frac{\lambda_n^k}{k!}\leq e^{-\lambda_n}\frac{\lambda_n^{\lfloor\lambda_n\rfloor}}{\lfloor\lambda_n\rfloor!}\leq n^{-\frac{1}{2}}. \tag{2.29}$$

We now use the fact that $n-k\leq T+1$ for $k\geq n-T$, $1-\frac{x}{j}\leq 1$, and (2.29) to conclude that the second summand in (2.28) is at most $\frac{(T+1)^2}{\prod_{j=1}^{T}(1-\frac{x}{j})n^{\frac{1}{2}}}$. Combining our lower and upper bounds on (2.28) proves the desired bound on the numerator of (2.27).

We now demonstrate that the denominator of (2.27) is bounded from below by

$$\prod_{j=1}^{T}(1-\frac{x}{j})\sum_{k=0}^{n-(T+1)}k\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\,e^{-\lambda_n}\frac{\lambda_n^k}{k!},$$

and bounded from above by

$$\prod_{j=1}^{T}(1-\frac{x}{j})(\sum_{k=0}^{n-(T+1)}k\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\,e^{-\lambda_n}\frac{\lambda_n^k}{k!}+\frac{(T+1)n^{\frac{1}{2}}}{\prod_{j=1}^{T}(1-\frac{x}{j})}).$$

Note that the denominator of (2.27) is equal to

$$\prod_{j=1}^{T}(1-\frac{x}{j})(\sum_{k=0}^{n-(T+1)} k \prod_{j=T+1}^{n-k}(1-\frac{x}{j})\, e^{-\lambda_n}\frac{\lambda_n^k}{k!} \tag{2.30}$$

$$+(\prod_{j=1}^{T}(1-\frac{x}{j}))^{-1}\sum_{k=n-T}^{n} k\prod_{j=1}^{n-k}(1-\frac{x}{j})\, e^{-\lambda_n}\frac{\lambda_n^k}{k!}).$$

We now bound the second summand in (2.30). Note that it is non-negative, yielding a trivial lower bound. Also, it is at most

$$\frac{T+1}{\prod_{j=1}^{T}(1-\frac{x}{j})}\max_{n-T\leq k\leq n} k\prod_{j=1}^{n-k}(1-\frac{x}{j})\, e^{-\lambda_n}\frac{\lambda_n^k}{k!} \leq \frac{(T+1)n^{\frac{1}{2}}}{\prod_{j=1}^{T}(1-\frac{x}{j})} \quad \text{by (2.29).}$$

Combining our lower and upper bounds on (2.30) proves the desired bound on the denominator of (2.27).

Combining the above upper and lower bounds for the numerator and denominator of (2.27), it follows that

$$\big(z_n(x)-\lambda_n\big)\lambda_n^{-\frac{1}{2}} \geq \lambda_n^{\frac{1}{2}}\frac{\sum_{k=0}^{n-(T+1)}(n-k)\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\, e^{-\lambda_n}\frac{\lambda_n^k}{k!}}{\sum_{k=0}^{n-(T+1)} k\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\, e^{-\lambda_n}\frac{\lambda_n^k}{k!}+\frac{(T+1)n^{\frac{1}{2}}}{\prod_{j=1}^{T}(1-\frac{x}{j})}}, \tag{2.31}$$

and

$$\big(z_n(x)-\lambda_n\big)\lambda_n^{-\frac{1}{2}} \leq \lambda_n^{\frac{1}{2}}\frac{\sum_{k=0}^{n-(T+1)}(n-k)\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\, e^{-\lambda_n}\frac{\lambda_n^k}{k!}+\frac{(T+1)^2}{\prod_{j=1}^{T}(1-\frac{x}{j})n^{\frac{1}{2}}}}{\sum_{k=0}^{n-(T+1)} k\prod_{j=T+1}^{n-k}(1-\frac{x}{j})\, e^{-\lambda_n}\frac{\lambda_n^k}{k!}}. \tag{2.32}$$

We now simplify the terms in (2.31) and (2.32) containing products of the form

$\prod_{j=T+1}^{n-k}(1-\frac{x}{j})$, by proving that for all $n \geq T+1$, and $k \in [0, n-T-1]$,

$$e^{-2T^{-1}}(n-k)^{-x}T^{x} \leq \prod_{j=T+1}^{n-k}(1-\frac{x}{j}) \leq e^{2T^{-1}}(n-k)^{-x}T^{x}. \tag{2.33}$$

Indeed, since $0 < x < 1$, by a simple Taylor series expansion we have that for $j \geq 3$, $1 \leq \frac{e^{-\frac{x}{j}}}{1-\frac{x}{j}} \leq 1+\frac{1}{j^2}$. Thus for $j \geq T+1$,

$$\prod_{j=T+1}^{n-k}\frac{e^{-\frac{x}{j}}}{(1-\frac{x}{j})} \leq \prod_{j=T+1}^{n-k}(1+\frac{1}{j^2}) \leq e^{\int_T^{\infty}\frac{1}{x^2}dx} = e^{T^{-1}},$$

and

$$e^{-T^{-1}}\prod_{j=T+1}^{n-k}e^{-\frac{x}{j}} \leq \prod_{j=T+1}^{n-k}(1-\frac{x}{j}) \leq \prod_{j=T+1}^{n-k}e^{-\frac{x}{j}}. \tag{2.34}$$

It follows from [118], and the fact that $n-k > T$, that

$$\log(\frac{n-k}{T}) - \frac{1}{2T} \leq H_{n-k} - H_T \leq \log(\frac{n-k}{T}) + \frac{1}{2T}. \tag{2.35}$$

Combining (2.34) and (2.35) yields

$$e^{-T^{-1}}e^{-\frac{x}{2T}}(n-k)^{-x}T^{x} \leq \prod_{j=T+1}^{n-k}(1-\frac{x}{j}) \leq e^{\frac{x}{2T}}(n-k)^{-x}T^{x}. \tag{2.36}$$

Since $0 < x < 1$ we have $0 < \frac{x}{2T} < \frac{1}{2T}$, and the desired bound follows from (2.36).

Combining (2.31), (2.32), and (2.33), with the assumptions on n, it follows that $\big(z_n(x) - \lambda_n\big)\lambda_n^{-\frac{1}{2}}$ is at least

$$\frac{e^{-2T^{-1}}\sum_{k=0}^{n-(T+1)}(n-k)^{1-x}e^{-\lambda_n}\frac{\lambda_n^k}{k!}}{e^{2T^{-1}}\sum_{k=0}^{n-(T+1)}k\lambda_n^{-\frac{1}{2}}(n-k)^{-x}\ e^{-\lambda_n}\frac{\lambda_n^k}{k!} + C_{T,x}}, \tag{2.37}$$

and at most

$$\frac{e^{2T^{-1}}\sum_{k=0}^{n-(T+1)}(n-k)^{1-x}\ e^{-\lambda_n}\frac{\lambda_n^k}{k!}+C_{T,x}n^{-\frac{1}{2}}}{e^{-2T^{-1}}\sum_{k=0}^{n-(T+1)}k\lambda_n^{-\frac{1}{2}}(n-k)^{-x}\ e^{-\lambda_n}\frac{\lambda_n^k}{k!}}. \tag{2.38}$$

With Inequalities (2.37) and (2.38) in hand, we are now in a position to complete the proof of Lemma 5. We begin by proving the lower bound. The term $\sum_{k=0}^{n-(T+1)}k\lambda_n^{-\frac{1}{2}}(n-k)^{-x}\ e^{-\lambda_n}\frac{\lambda_n^k}{k!}$ appearing in the denominator of (2.37) is at most

$$\sum_{k=0}^{\lceil n-T^{-1}n^{\frac{1}{2}}\rceil} k(n-k)^{-x}\lambda_n^{-\frac{1}{2}}\frac{e^{-\lambda_n}\lambda_n^k}{k!} \tag{2.39}$$

$$+\max_{\lceil n-T^{-1}n^{\frac{1}{2}}\rceil+1\le k\le n}\left(k\lambda_n^{-\frac{1}{2}}\ e^{-\lambda_n}\frac{\lambda_n^k}{k!}\right)\sum_{k=\lceil n-T^{-1}n^{\frac{1}{2}}\rceil+1}^{n-(T+1)}(n-k)^{-x}. \tag{2.40}$$

It follows from (2.29), and the fact that $\frac{n}{\lambda_n}\le 2$, that the second summand of (2.39) is at most

$$\frac{n}{\lambda_n^{\frac{1}{2}}}n^{-\frac{1}{2}}\sum_{k=\lceil n-T^{-1}n^{\frac{1}{2}}\rceil+1}^{n-(T+1)}(n-k)^{-x}\ \le\ (\frac{n}{\lambda_n})^{\frac{1}{2}}\int_0^{T^{-1}n^{\frac{1}{2}}}y^{-x}dy\ \le\ \frac{2T^{-(1-x)}n^{\frac{1-x}{2}}}{1-x}.$$

Using the above to upper-bound the denominator of (2.37), and then multiplying both numerator and denominator by $\lambda_n^{\frac{x-1}{2}}$, completes the proof of the lower bound.

We now prove the upper bound. By non-negativity the denominator of (2.38) is at least

$$e^{-2T^{-1}}\sum_{k=0}^{\lceil n-T^{-1}n^{\frac{1}{2}}\rceil}k\lambda_n^{-\frac{1}{2}}(n-k)^{-x}\frac{e^{-\lambda_n}\lambda_n^k}{k!}.$$

The upper bound then follows from (2.38) and multiplying both numerator and denominator by $\lambda_n^{\frac{x-1}{2}}$, completing the proof of the lemma. □

Chapter 3

Explicit Bounds on the Distance to Steady-state for the M/M/n Queue in the Halfin-Whitt Regime

3.1 Introduction and Literature Review

It is well-known that the steady-state behavior of the $M/M/n$ queue in the H-W regime is quite simple in practice [52], while the transient dynamics are more complicated [52], and it is common to use the steady-state approximation to the transient distribution. Thus it is important to understand the quality of the steady-state approximation. In Chapter 2, we identified the rate of convergence to steady-state of the $M/M/n$ queue in the H-W regime, up to exponential order. However, in many applications it is desirable to have explicit bounds on the error of the steady-state approximation, as opposed to just an understanding of its behavior up to exponential order. It seems that prior to this thesis, no such explicit bounds had been studied for the $M/M/n$ queue in the H-W regime.

As discussed in Section 2.1, the question of how quickly the positive recurrent $M/M/n$ queue approaches stationarity has a rich history in the queueing literature. The most relevant work is that of Karlin and McGregor [62], who worked out a powerful and elegant theory that gave explicit integral representations for the transient distributions of b-d-p. In [63], they applied their framework to the $M/M/n$ queue, giving integral representations for the transient distribution of the $M/M/n$ queue. As explained in Section 2.1, these representations involve integrals w.r.t. a certain spectral measure, which is itself defined in terms of the set of zeros of a high-degree polynomial. Thus these integrals may be difficult to evaluate numerically, and it is unclear how to translate the representations into simple, usable bounds in the H-W regime.

Although much of the literature on the rate of convergence to steady-state of the $M/M/n$ queue has focused on the exponential rate of convergence (i.e. spectral gap, see Section 2.1), some explicit bounds have been proven. In [119], Zeifman used tools from the theory of differential equations to give explicit bounds on the total variational distance between the transient and steady-state distributions of a b-d-p, and explicitly examines the $M/M/n$ queue. In [106], van Doorn and Zeifman used the techniques developed in [119] to derive explicit bounds on the distance to stationarity for a different queueing model, and examined how their bounds perform in a certain heavy-traffic regime (not H-W). In [107], the authors extended the results of [119] and [106]. In [23], Chen developed very general bounds for the distance to stationarity for Markov chains, and then applied these to b-d-p. However, these bounds are generally not studied in the H-W regime, and thus may not scale desirably with n in the H-W regime. We note that the complexity of bounding the distance to stationarity uniformly for a sequence of b-d-p is related to the cutoff phenomenon for Markov chains [31], which has been studied in the context of queueing systems [42].

In this chapter, we prove explicit bounds on the distance to stationarity for the

$M/M/n$ queue in the H-W regime, when $B < B^*$, e.g. characterizing the error in estimating the transient probability of delay by the corresponding steady-state quantity. Our bounds hold for any sufficiently large fixed n, i.e. number of servers, and scale independently of n in the H-W regime. Also, we use our bounds to provide a heuristic rule-of-thumb which could be used to determine the time it takes an overloaded (underloaded) queueing system to return (probabilistically) to the steady-state.

3.1.1 Outline of chapter

The rest of the chapter proceeds as follows. In Section 3.2, we state our main results. In Section 3.3, we present the proof of our main theorem. In Section 3.4, we compare our bounds to other bounds from the literature. In Section 3.5 we summarize our main results and present ideas for future research. We include a technical appendix in Section 3.6.

3.2 Main Results

3.2.1 Definitions and notations

We now recall several important quantities (defined in Chapter 2) for the $M/M/n$ queue Q^n, namely the $M/M/n$ queue with arrival rate $\lambda_n = n - Bn^{\frac{1}{2}}$ and service rate $\mu = 1$. Recall that $Q^n(t)$ denotes the number in system at time t in Q^n. As in Chapter 2, $P^n_{i,j}(t) = \Pr\big(Q^n(t) = j | Q^n(0) = i\big)$, $P^n_j(\infty) = \Pr(Q^n(\infty) = j)$, $P^n_{i,\leq j}(t) = \sum_{k=0}^{j} P^n_{i,k}(t)$, and $P^n_{\leq j}(\infty) = \sum_{k=0}^{j} P^n_k(\infty)$. Unless otherwise stated, all functions are defined only for real values of x. All empty products are assumed to be equal to unity, and all empty summations are assumed to be equal to zero.

3.2.2 Main results

We now state our explicit bounds on the distance to stationarity for the case $B < B^*$.

Theorem 9. *Let us fix some $B \in (0, B^*)$ and $a_1, a_2 \in \mathbb{R}$. Let $\alpha = \max(|a_1|, |a_2|)$. Then there exists $N_{B,a_1,a_2} < \infty$, depending only on B, a_1, a_2, s.t. for all $n \geq N_{B,a_1,a_2}$ and $t \geq 1$,*

$$|n^{\frac{1}{2}} P^n_{\lceil n+a_1 n^{\frac{1}{2}} \rceil, \lceil n+a_2 n^{\frac{1}{2}} \rceil}(t) - n^{\frac{1}{2}} P^n_{\lceil n+a_2 n^{\frac{1}{2}} \rceil}(\infty)| \leq t^{-\frac{1}{2}} \exp\big(120(\alpha^2+1) - \frac{B^2}{4}t\big), \quad (3.1)$$

and

$$|P^n_{\lceil n+a_1 n^{\frac{1}{2}} \rceil, \leq \lceil n+a_2 n^{\frac{1}{2}} \rceil}(t) - P^n_{\leq \lceil n+a_2 n^{\frac{1}{2}} \rceil}(\infty)| \leq B^{-1} t^{-\frac{1}{2}} \exp\big(120(\alpha^2+1) - \frac{B^2}{4}t\big). \quad (3.2)$$

Note that Theorem 9 provides a bound for any sufficiently large *fixed* n and *all times* greater than 1, which is stronger than the bounds that would follow by naively applying the weak-convergence theory. Indeed, using the weak-convergence theory one could derive bounds of the form 'for any *fixed* time t, there exists a sufficiently large N_t s.t. $n \geq N_t$ implies ...', but these bounds would be too weak to make statements about *all time* for any *fixed* n. Furthermore, in light of the results of Chapter 2, the exponent $\frac{B^2}{4}$ is asymptotically the best possible. We note that although we were able to derive partial results for the case $B \geq B^*$, the derived bounds were considerably more complicated than those of Theorem 9, and we leave it as an open question to derive simple explicit bounds for the case $B \geq B^*$.

We now discuss a practical 'rule-of-thumb' interpretation of Theorem 9 in the context of managing call centers. Suppose one is operating a call center in the H-W regime (with $0 < B < B^*$) and at time $t = 0$ the center begins either very overloaded ($Q(0) = n + an^{\frac{1}{2}}$) or underloaded ($Q(0) = n - an^{\frac{1}{2}}$) with $a \gg 0$ a large constant (independent of n). Then Theorem 9 says that for any $\epsilon > 0$, the probability that a caller at time $t^*_{a,B} \triangleq \frac{240}{B^2}a^2 + \frac{4}{B^2}log(\frac{1}{\epsilon})$ has to wait for service will be roughly within ϵ of the probability that a caller in steady-state has to wait for service. In particular, after $t^*_{a,B}$ time units have elapsed, the system will have returned to within ϵ of steady-state.

3.3 Proof of Main Result

In this section we prove our main result.

3.3.1 K-M representation when the traffic intensity is at least ρ^*_n.

In this subsection we formally state Karlin and McGregor's representation for the transient distribution of the $M/M/n$ queue, as derived in [63] and reviewed in [102], when the traffic intensity is at least ρ^*_n (defined in Section 2.1). In particular, recall from Section 2.1, in particular Theorem 3, that in [102] van Doorn proved that for any fixed n, there exists $\rho^*_n \in [0, 1)$ s.t. if $\frac{\lambda_n}{n} \geq \rho^*_n$, then the spectral measure appearing in the K-M integral representation for $\{P^n_{i,j}(t), i, j \geq 0\}$ has no jumps away from the origin. As proven in [102], in this case the K-M integral representation simplifies considerably. We now introduce some additional notations, and then explicitly give the K-M integral representation for $P^n_{i,j}(t)$, for the case $\frac{\lambda_n}{n} \geq \rho^*_n$. We define

$$Q_{n,k}(x) \triangleq \begin{cases} 1 & \text{if k = 0;} \\ 1 - \frac{x}{\lambda_n} & \text{if k = 1;} \\ (1 - \frac{x}{\lambda_n} + \frac{\min(k-1,n)}{\lambda_n})Q_{n,k-1}(x) - \frac{\min(k-1,n)}{\lambda_n}Q_{n,k-2}(x) & \text{otherwise;} \end{cases} \qquad (3.3)$$

and let $c_n(x) \triangleq Q_{n,n}^2(x) - Q_{n,n-1}(x)Q_{n,n+1}(x)$. We note that the $Q_{n,k}(x)$ polynomials can be related to the well-studied Poisson-Charlier polynomials, and refer the reader to [102] for details. Also, it is proven in [102] that $x \in \big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2, (n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2\big)$ implies $c_n(x) > 0$. We define

$$b_n(x) \triangleq \begin{cases} \big(4\lambda_n n - (\lambda_n + n - x)^2\big)^{\frac{1}{2}} & \text{if } (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 \le x \le (n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2; \\ \infty & \text{otherwise.} \end{cases}$$

Note that $b_n(x)$ is real-valued for $x \in [(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2, (n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2]$, and therefore

$$b_n(x) = \big(x - (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)^{\frac{1}{2}}\big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - x\big)^{\frac{1}{2}}. \tag{3.4}$$

We also define

$$g_n(k) \triangleq \begin{cases} \frac{\lambda_n^k}{k!} & \text{if } \ 0 \le k \le n; \\ \frac{\lambda_n^n}{n!}\left(\frac{\lambda_n}{n}\right)^{k-n} & \text{otherwise.} \end{cases}$$

We now formally state Karlin and McGregor's representation for the transient distribution of the $M/M/n$ queue, as derived in [63] and reviewed in [102], when the traffic intensity is at least ρ_n^*. Namely, it follows from the K-M representation that if $\frac{\lambda_n}{n} \ge \rho_n^*$, then for all $i, j \ge 0$,

Theorem 10.

$$P_{i,j}^n(t) - P_j^n(\infty) = (2\pi)^{-1} g_n(j) \frac{n!}{\lambda_n^n} (\lambda_n n)^{-1} \int_{(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2} e^{-xt} Q_{n,i}(x) Q_{n,j}(x) b_n(x) c_n(x)^{-1} dx.$$

3.3.2 Bounds for $|Q_{n,n}(x)|$, $|Q_{n,n-1}(x)|$, and $|Q_{n,n}(x) - Q_{n,n-1}(x)|$

In this subsection we prove some bounds for $|Q_{n,n}(x)|$, $|Q_{n,n-1}(x)|$, and $|Q_{n,n}(x) - Q_{n,n-1}(x)|$ in terms of $c_n(x)$ and $b_n(x)$. Let $h_n(x) \triangleq 2nb_n(x)^{-1}$. Then we prove that

Lemma 17. *For all* $x \in \big((n^{\frac{1}{2}} - \sqrt{\lambda_n})^2, (n^{\frac{1}{2}} + \sqrt{\lambda_n})^2\big)$,

(i) $|Q_{n,n}(x)|c_n(x)^{-\frac{1}{2}} \leq h_n(x)$,

(ii) $|Q_{n,n-1}(x)|c_n(x)^{-\frac{1}{2}} \leq h_n(x)$,

(iii) $|Q_{n,n}(x) - Q_{n,n-1}(x)|c_n(x)^{-\frac{1}{2}} \leq (\frac{x}{n})^{\frac{1}{2}} h_n(x)$.

Proof. It follows from the definition of the $Q_{n,k}(x)$ polynomials that

$$c_n(x) = Q^2_{n,n}(x) - \frac{\lambda_n + n - x}{\lambda_n} Q_{n,n}(x)Q_{n,n-1}(x) + \frac{n}{\lambda_n} Q^2_{n,n-1}(x). \tag{3.5}$$

We now prove assertion i. First, note that $b_n(x) > 0$ and $c_n(x) = Q^2_{n,n}(x) - Q_{n,n-1}(x)Q_{n,n+1}(x) > 0$, ensuring that $c_n(x)^{-1}$ and $b_n(x)^{-1}$ are well-defined. If $Q_{n,n}(x) = 0$, then $|Q_{n,n}(x)|c_n(x)^{-\frac{1}{2}} = 0 < h_n(x)$. If $Q_{n,n}(x) \neq 0$ then

$$\begin{aligned} Q^2_{n,n}(x)c_n(x)^{-1} &= \Big(1 - \frac{\lambda_n + n - x}{\lambda_n}\frac{Q_{n,n-1}(x)}{Q_{n,n}(x)} + \frac{n}{\lambda_n}\big(\frac{Q_{n,n-1}(x)}{Q_{n,n}(x)}\big)^2\Big)^{-1} \\ &\leq \sup_{z\in\mathbb{R}} \big(1 - \frac{\lambda_n + n - x}{\lambda_n} z + \frac{n}{\lambda_n} z^2\big)^{-1}. \end{aligned} \tag{3.6}$$

By elementary calculus, $\alpha_1(z) \triangleq 1 - \frac{\lambda_n + n - x}{\lambda_n} z + \frac{n}{\lambda_n} z^2$ is a convex function of z minimized at $z = \frac{\lambda_n + n - x}{2n}$, and $\alpha_1(\frac{\lambda_n+n-x}{2n}) = \frac{b^2_n(x)}{4\lambda_n n}$. Combining with (3.6), taking the square root of both sides, and observing that $2\sqrt{\lambda_n n} < 2n$ completes the proof of i.

We now prove assertion ii. If $Q_{n,n-1}(x) = 0$, then $|Q_{n,n-1}(x)|c_n(x)^{-\frac{1}{2}} = 0 < h_n(x)$.

If $Q_{n,n-1}(x) \neq 0$ then

$$\begin{aligned} Q^2_{n,n-1}(x)c_n(x)^{-1} &= \big((\frac{Q_{n,n}(x)}{Q_{n,n-1}(x)})^2 - \frac{\lambda_n + n - x}{\lambda_n}\frac{Q_{n,n}(x)}{Q_{n,n-1}(x)} + \frac{n}{\lambda_n}\big)^{-1} \\ &\leq \sup_{z\in\mathbb{R}} \big(z^2 - \frac{\lambda_n + n - x}{\lambda_n} z + \frac{n}{\lambda_n}\big)^{-1}. \end{aligned} \tag{3.7}$$

Letting $\alpha_2(z) \triangleq z^2 - \frac{\lambda_n+n-x}{\lambda_n}z + \frac{n}{\lambda_n}$, we find by elementary calculus that $\alpha_2(z)$ is a convex function of z minimized at $z = \frac{\lambda_n+n-x}{2\lambda_n}$, and $\alpha_2(\frac{\lambda_n+n-x}{2\lambda_n}) = \frac{b_n^2(x)}{4\lambda_n^2}$. Combining with (3.7), taking the square root of both sides, and observing that $2\lambda_n < 2n$ completes the proof of ii.

We now prove assertion iii. It is shown in [102] that $Q_{n,n}(x)$ and $Q_{n,n-1}(x)$ do not have any common zeros. Thus first suppose $Q_{n,n}(x) = 0$. Then from (3.5), $c_n(x) = \frac{n}{\lambda_n}Q^2_{n,n-1}(x)$. Thus

$$\big(Q_{n,n}(x) - Q_{n,n-1}(x)\big)^2 c_n(x)^{-1} = \frac{Q^2_{n,n-1}(x)}{\frac{n}{\lambda_n}Q^2_{n,n-1}(x)} = \frac{\lambda_n}{n}. \tag{3.8}$$

Furthermore,

$$\frac{4\lambda_n x}{b_n^2(x)} = \frac{4\lambda_n x}{4\lambda_n n - (\lambda_n + n - x)^2} = 1 + \frac{(\lambda_n + x - n)^2}{b_n^2(x)} \geq 1. \tag{3.9}$$

Combining (3.8) and (3.9) with the fact that $\frac{\lambda_n}{n} < 1$, we find that

$$\big(Q_{n,n}(x) - Q_{n,n-1}(x)\big)^2 c_n(x)^{-1} \leq \frac{4\lambda_n x}{b_n^2(x)}. \tag{3.10}$$

Now suppose $Q_{n,n-1}(x) = 0$. Then from (3.5), $c_n(x) = Q^2_{n,n}(x)$, and it follows from (3.9) that

$$\big(Q_{n,n}(x) - Q_{n,n-1}(x)\big)^2 c_n(x)^{-1} = \frac{Q^2_{n,n}(x)}{Q^2_{n,n}(x)} = 1 \leq \frac{4\lambda_n x}{b_n^2(x)}. \tag{3.11}$$

Now suppose $Q_{n,n}(x) \neq 0$ and $Q_{n,n-1}(x) \neq 0$. Then

$$\begin{aligned}\big(Q_{n,n}(x) - Q_{n,n-1}(x)\big)^2 c_n(x)^{-1} &= \frac{(\frac{Q_{n,n}(x)}{Q_{n,n-1}(x)} - 1)^2}{(\frac{Q_{n,n}(x)}{Q_{n,n-1}(x)})^2 - \frac{\lambda_n+n-x}{\lambda_n}\frac{Q_{n,n}(x)}{Q_{n,n-1}(x)} + \frac{n}{\lambda_n}} \\ &\leq \sup_{z\in\mathbb{R}} \frac{(z-1)^2}{z^2 - \frac{\lambda_n+n-x}{\lambda_n}z + \frac{n}{\lambda_n}}. \qquad (3.12)\end{aligned}$$

Let $\alpha_3(z) \triangleq \frac{(z-1)^2}{z^2 - \frac{\lambda_n+n-x}{\lambda_n}z + \frac{n}{\lambda_n}}$. Note that $\inf_{z\in\mathbb{R}}(z^2 - \frac{\lambda_n+n-x}{\lambda_n}z + \frac{n}{\lambda_n}) > 0$ from our earlier analysis of $\alpha_2(z)$. Thus by elementary calculus, $\alpha_3(z)$ is a continuously differentiable rational function of z on $\mathbb{R}$, and

$$\frac{d}{dz}\alpha_3(z) = \frac{(z-1)\big((n-\lambda_n+x) + (\lambda_n - n + x)z\big)}{\lambda_n(z^2 - \frac{\lambda_n+n-x}{\lambda_n}z + \frac{n}{\lambda_n})^2}.$$

It follows that the zeros of $\frac{d}{dz}\alpha_3(z)$ occur at $z = 1$ and $z = \frac{\lambda_n-n-x}{\lambda_n-n+x}$. Thus by elementary calculus, $\sup_{z\in\mathbb{R}}\alpha_3(z)$ must be one of $\alpha_3(1), \alpha_3(\frac{\lambda_n-n-x}{\lambda_n-n+x}), \lim_{z\to-\infty}\alpha_3(z), \lim_{z\to\infty}\alpha_3(z)$. Trivially, $\alpha_3(1) = 0$, and $\lim_{z\to-\infty}\alpha_3(z) = \lim_{z\to\infty}\alpha_3(z) = 1$. Furthermore,

$$\alpha_3(\frac{\lambda_n-n-x}{\lambda_n-n+x}) = \frac{(\frac{2x}{\lambda_n-n+x})^2}{(1-\frac{2x}{\lambda_n-n+x})^2 - \frac{\lambda_n+n-x}{\lambda_n}(1-\frac{2x}{\lambda_n-n+x}) + \frac{n}{\lambda_n}} = \frac{4x\lambda_n}{b_n^2(x)}.$$

Combining the above with (3.9), we find that $\sup_{z\in\mathbb{R}}\alpha_3(z) \leq \frac{4\lambda_n x}{b_n^2(x)}$. Combining with (3.10) - (3.12) and taking square roots completes the proof of the lemma. □

3.3.3 Bounding $Q_{n,k}(x)$ for $k = n \pm O(n^{\frac{1}{2}})$

In this subsection, we bound the $Q_{n,k}(x)$ polynomials for $k = n \pm O(n^{\frac{1}{2}})$. In particular, we prove that

Lemma 18. *For each $\alpha \geq B > 0$, there exists N_α s.t. $n \geq N_\alpha$, $x \in \big((n^{\frac{1}{2}} - \sqrt{\lambda_n})^2, (n^{\frac{1}{2}} + \sqrt{\lambda_n})^2\big)$, and $k \in [n - \alpha n^{\frac{1}{2}} - 2, n + \alpha n^{\frac{1}{2}} + 2]$ implies that* *(i)* $|Q_{n,k}(x)|c_n(x)^{-\frac{1}{2}} \le h_n(x)\exp\big((\alpha + n^{-\frac{1}{5}})(\alpha + 3x^{\frac{1}{2}})\big)$,

(ii) $|Q_{n,k+1}(x) - Q_{n,k}(x)|c_n(x)^{-\frac{1}{2}} \le h_n(x)r_n(x)^{\frac{1}{2}}\exp\big((\alpha + n^{-\frac{1}{5}})(\alpha + 3x^{\frac{1}{2}})\big)$.

We proceed by bounding the growth of $|Q_{n,k+1}(x)|$ and $|Q_{n,k+1}(x) - Q_{n,k}(x)|$ as k diverges from n, and then combining with Lemma 17. We begin by proving some simple asymptotic bounds. We define $q_n \triangleq 1 + n^{-\frac{1}{4}}$, $r_n(x) \triangleq q_n x n^{-1}$, and $s_n(\alpha) \triangleq q_n \alpha n^{-\frac{1}{2}}$. Then

Lemma 19. *For each $\alpha \ge B > 0$, there exists N_α s.t. $n \ge N_\alpha$, $k \in [n - \alpha n^{\frac{1}{2}} - 3, n + \alpha n^{\frac{1}{2}} + 3]$, and $x > 0$ implies $|\frac{x}{\min(k,n)}|, |\frac{x}{\lambda_n}| \le r_n(x)$; and $|\frac{\lambda_n}{\min(k,n)}|, |\frac{n}{\lambda_n}| \le 1 + s_n(\alpha)$.*

Proof. For any constants $a > 0$ and $b \ge 0$ and all sufficiently large n, $n\big(n - an^{\frac{1}{2}} - b\big)^{-1} \le 1 + s_n(a)$, and $(n - an^{\frac{1}{2}} - b)^{-1} \le q_n n^{-1}$. The lemma then follows from a simple case analysis. □

We now bound the growth of $|Q_{n,k+1}(x)|$ and $|Q_{n,k+1}(x) - Q_{n,k}(x)|$ as k diverges from n, proving that

Lemma 20. *For each $\alpha \ge B > 0$, there exists N_α s.t. $n \ge N_\alpha$, $k \in [n - \alpha n^{\frac{1}{2}} - 3, n + \alpha n^{\frac{1}{2}} + 3]$, $x > 0$, and $i \in \{1, -1\}$ implies that*

(i) $|Q_{n,k+i}(x)| \le \big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(|Q_{n,k}(x)| + |Q_{n,k}(x) - Q_{n,k-i}(x)|\big)$,

(ii) $|Q_{n,k+i}(x) - Q_{n,k}(x)| \le \big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(r_n(x)|Q_{n,k}(x)| + |Q_{n,k}(x) - Q_{n,k-i}(x)|\big)$.

Proof. Let us choose N_α to satisfy the conditions of Lemma 19 (the existence of such an N_α follows from the same lemma), and suppose $n \ge N_\alpha$. We first treat the case $i = 1$, and begin by proving assertion i. Recall that for $k \ge 1$, $|Q_{n,k+1}(x)|$ equals $\big|(1 - \frac{x}{\lambda_n} + \frac{\min(k,n)}{\lambda_n})Q_{n,k}(x) - \frac{\min(k,n)}{\lambda_n}Q_{n,k-1}(x)\big|$, which is at most

$$\big(1 + r_n(x)\big)|Q_{n,k}(x)| + \big(1 + s_n(\alpha)\big)|Q_{n,k}(x) - Q_{n,k-1}(x)|, \tag{3.13}$$

by the triangle inequality and Lemma 19. Assertion i. follows by multiplying the first summand of (3.13) by $1 + s_n(\alpha)$, and the second summand by $1 + r_n(x)$.

We now prove assertion ii. for the case $i = 1$. Rearranging Definition (3.3), we find that for $k \geq 1$, $|Q_{n,k+1}(x) - Q_{n,k}(x)|$ is equal to $|-\frac{x}{\lambda_n} Q_{n,k}(x) + \frac{\min(k,n)}{\lambda_n}\big(Q_{n,k}(x) - Q_{n,k-1}(x)\big)|$, which is at most

$$r_n(x)|Q_{n,k}(x)| + \big(1 + s_n(\alpha)\big)|Q_{n,k}(x) - Q_{n,k-1}(x)|, \tag{3.14}$$

by the triangle inequality and Lemma 19. Assertion ii. then follows by multiplying the first summand of (3.14) by $\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)$, and the second summand by $\big(1 + r_n(x)\big)$.

The proofs for the case $i = -1$ follow very similarly, and are omitted. □

We now use Lemma 20 to bound $|Q_{n,k}(x)|$ for values of k that are $O(n^{\frac{1}{2}})$ away from n.

Lemma 21. *For each $\alpha \geq B > 0$, there exists N_α s.t. $n \geq N_\alpha$, $k \in [n - \alpha n^{\frac{1}{2}} - 3, n + \alpha n^{\frac{1}{2}} + 3]$, and $x \in \big((n^{\frac{1}{2}} - \sqrt{\lambda_n})^2, (n^{\frac{1}{2}} + \sqrt{\lambda_n})^2\big)$ implies that*

(i) $|Q_{n,k}(x)| \leq \Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{|k-n|} c_n(x)^{\frac{1}{2}} h_n(x)$,

(ii) $|Q_{n,k}(x) - Q_{n,k+I(k<n)-I(k\geq n)}(x)|$ *is at most*

$$\Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{|k-n|} r_n(x)^{\frac{1}{2}} c_n(x)^{\frac{1}{2}} h_n(x).$$

Proof. Let us choose N_α to satisfy the conditions of Lemma 20 (the existence of such an N_α follows from the same lemma), and suppose $n \geq N_\alpha$. We first treat the case $k \geq n$. We proceed by induction on assertions i. and ii. simultaneously. The base case $k = n$ follows immediately from Lemma 17.i and Lemma 17.iii. Now, suppose the induction is true for some $k \in [n, n + \alpha n^{\frac{1}{2}} + 2]$. Then by Lemma 20, $|Q_{n,k+1}(x)|$ is at most

$$\begin{aligned}
& \big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(|Q_{n,k}(x)| + |Q_{n,k}(x) - Q_{n,k-1}(x)|\big) \\
\leq \; & \big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\Bigg(\Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{k-n} c_n(x)^{\frac{1}{2}} h_n(x) \\
+ \; & \Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{k-n} r_n(x)^{\frac{1}{2}} c_n(x)^{\frac{1}{2}} h_n(x)\Bigg) \text{ by ind. hyp.} \\
= \; & \Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{k+1-n} c_n(x)^{\frac{1}{2}} h_n(x). \qquad (3.15)
\end{aligned}$$

Similarly, by Lemma 20, $||Q_{n,k+1}(x) - Q_{n,k}(x)|$ is at most

$$\begin{aligned}
& \big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(r_n(x)|Q_{n,k}(x)| + |Q_{n,k}(x) - Q_{n,k-1}(x)|\big) \\
\leq \; & \big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\Bigg(\Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{k-n} r_n(x) c_n(x)^{\frac{1}{2}} h_n(x) \\
+ \; & \Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{k-n} r_n(x)^{\frac{1}{2}} c_n(x)^{\frac{1}{2}} h_n(x)\Bigg) \text{ by ind. hyp.} \\
= \; & \Big(\big(1 + r_n(x)\big)\big(1 + s_n(\alpha)\big)\big(1 + r_n(x)^{\frac{1}{2}}\big)\Big)^{k+1-n} r_n(x)^{\frac{1}{2}} c_n(x)^{\frac{1}{2}} h_n(x). \qquad (3.16)
\end{aligned}$$

This concludes the induction, proving assertions i. and ii. for the case $k \geq n$.

The proof for the case $k < n$ follows from a very similar argument, and is omitted.

□

We now complete the proof of Lemma 18.

Proof of Lemma 18. By Lemma 21, we find that there exists N_α s.t. $n \geq N_\alpha$ implies

$$\begin{aligned} |Q_{n,k}(x)|c_n(x)^{-\frac{1}{2}} &\leq h_n(x)\exp\big((\alpha n^{\frac{1}{2}} + 2)(r_n(x) + s_n(\alpha) + r_n(x)^{\frac{1}{2}})\big) \\ &= h_n(x)\exp\big((\alpha n^{\frac{1}{2}} + 2)(q_n x n^{-1} + q_n \alpha n^{-\frac{1}{2}} + q_n^{\frac{1}{2}} x^{\frac{1}{2}} n^{-\frac{1}{2}})\big) \\ &\leq h_n(x)\exp\big(q_n(\alpha + 2n^{-\frac{1}{2}})(xn^{-\frac{1}{2}} + \alpha + x^{\frac{1}{2}})\big). \qquad (3.17) \end{aligned}$$

By identical reasoning,

$$|Q_{n,k+1}(x) - Q_{n,k}(x)|c_n(x)^{-\frac{1}{2}} \leq h_n(x)r_n(x)^{\frac{1}{2}}\exp\big(q_n(\alpha + 2n^{-\frac{1}{2}})(xn^{-\frac{1}{2}} + \alpha + x^{\frac{1}{2}})\big). \qquad (3.18)$$

Furthermore, note that since

$$\frac{xn^{-\frac{1}{2}}}{2x^{\frac{1}{2}}} = \frac{x^{\frac{1}{2}}}{2n^{\frac{1}{2}}} < \frac{n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}}}{2n^{\frac{1}{2}}} < 1,$$

it will be the case that $xn^{-\frac{1}{2}} < 2x^{\frac{1}{2}}$ for $x \in \big(0, (n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2\big)$. The lemma follows by combining with (3.17), (3.18), and the fact that $q_n(\alpha + 2n^{-\frac{1}{2}}) \leq \alpha + n^{-\frac{1}{5}}$ for all sufficiently large n. □

3.3.4 Proof of Theorem 9

In this subsection we complete the proof of our main result, Theorem 9. We begin by deriving a variant of Theorem 10 for $P^n_{i,\leq j}(t)$, as opposed to $P^n_{i,j}(t)$, that does not simply sum over all $j+1$ states $\leq j$, but instead relies on a 'probability flow' interpretation using the Chapman-Kolmogorov equation. This will allow us to bound the distance to stationarity uniformly in n for the c.d.f., since we can sidestep having to sum over 'too many' error terms.

Lemma 22. *For $\frac{\lambda_n}{n} \geq \rho_n^*$, $|P^n_{i,\leq j}(t) - P^n_{\leq j}(\infty)|$ is at most*

$$(2\pi)^{-1} g_n(j) \frac{n!}{\lambda_n^n} n^{-1} \int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} e^{-xt} x^{-1} |Q_{n,i}(x)||Q_{n,j+1}(x) - Q_{n,j}(x)| b_n(x) c_n(x)^{-1} dx.$$

Proof. It follows from the Chapman-Kolmogorov equation and a straightforward telescoping sum argument that $\frac{d}{dt}P^n_{i,\leq j}(t) = \min(j+1,n)P^n_{i,j+1}(t) - \lambda_n P^n_{i,j}(t)$. Thus for all $i, t \geq 0$, $P^n_{\leq j}(\infty) = P^n_{i,\leq j}(t) + \int_t^\infty \big(\min(j+1,n)P^n_{i,j+1}(s) - \lambda_n P^n_{i,j}(s)\big)ds$. It follows that $|P^n_{i,\leq j}(t) - P^n_{\leq j}(\infty)| = \big|\int_t^\infty \big(\min(j+1,n)P^n_{i,j+1}(s) - \lambda_n P^n_{i,j}(s)\big)ds\big|$, which by detailed balance equals

$$\Big|\int_t^\infty \Big(\min(j+1,n)\big(P^n_{i,j+1}(s) - P^n_{i,j+1}(\infty)\big) - \lambda_n\big(P^n_{i,j}(s) - P^n_{i,j}(\infty)\big)\Big)ds\Big|$$

$$= \Big|\int_t^\infty \Big((2\pi)^{-1} g_n(j)\frac{n!}{\lambda_n^n} n^{-1} \int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} e^{-xs} Q_{n,i}(x)\big(Q_{n,j+1}(x) - Q_{n,j}(x)\big) b_n(x) c_n(x)^{-1} dx\Big) ds\Big|,$$

by Theorem 10, since $\min(j+1,n)g_n(j+1)\frac{n!}{\lambda_n^n}(\lambda_n n)^{-1} = g_n(j)\frac{n!}{\lambda_n^n}n^{-1}$. Further simplifying, we find the above to be at most

$$\int_t^\infty \Big((2\pi)^{-1} g_n(j)\frac{n!}{\lambda_n^n} n^{-1} \int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} e^{-xs} |Q_{n,i}(x)||Q_{n,j+1}(x) - Q_{n,j}(x)| b_n(x) c_n(x)^{-1} dx\Big) ds$$

$$= (2\pi)^{-1} g_n(j)\frac{n!}{\lambda_n^n} n^{-1} \int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} e^{-xt} x^{-1} |Q_{n,i}(x)||Q_{n,j+1}(x) - Q_{n,j}(x)| b_n(x) c_n(x)^{-1} dx$$

by Tonelli's Theorem, and the lemma follows. □

We now prove bounds on a special type of integral that arises in the analysis of $P^n_{i,j}(t) - P^n_j(\infty)$.

Lemma 23. *For all $B, C > 0$ there exists $N_{B,C}$ s.t. $n \geq N_{B,C}$ implies that for all*

$c \in (0, C]$ *and* $t \geq 1$,

$$\int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} \exp\big(-xt+cx^{\frac{1}{2}}\big)b_n(x)^{-1}dx \leq q_n(\frac{\pi}{t\lambda_n})^{\frac{1}{2}} \exp\big(\frac{1}{2}Bc+c^2(2t)^{-1}-\frac{B^2}{4}t\big). \quad (3.19)$$

Proof. The proof is deferred to the appendix. □

Before completing the proof of our main result, we derive an asymptotic bound for $g_n(k)$, proving that

Lemma 24. *For each* $B > 0$, *there exists* N_B *s.t.* $n \geq N_B$ *implies* $\sup_{k\geq 0} g_n(k)\frac{n!}{\lambda_n^n} \leq q_n \exp(B^2)$.

Proof. Since $\frac{\lambda_n}{n} < 1$, we have that $g_n(k) \leq \frac{\lambda_n^{\lfloor \lambda_n \rfloor}}{\lfloor \lambda_n \rfloor !}$ for all $k \geq 0$. Thus

$$g_n(k)\frac{n!}{\lambda_n^n} \leq \frac{\lambda_n^{\lfloor \lambda_n \rfloor}}{\lfloor \lambda_n \rfloor !}\frac{n!}{\lambda_n^n} \leq (\frac{n}{\lambda_n})^{n-\lfloor \lambda_n \rfloor} = \exp\big(B^2 + O(n^{-\frac{1}{2}})\big),$$

and the lemma follows from a simple Taylor series expansion. □

We now complete the proof of Theorem 9.

Proof of Theorem 9. Suppose $B \in (0, B^*)$, $a_1, a_2 \in \mathbb{R}$. Let $\alpha = \max(B, |a_1|, |a_2|)$, $i = \lceil n + a_1 n^{\frac{1}{2}} \rceil$, and $j = \lceil n + a_2 n^{\frac{1}{2}} \rceil$. Then it follows from Theorem 10, Corollary 1, Lemma 24, and Lemma 18 that for all sufficiently large n and all $t \geq 1$, the l.h.s. of (3.1) is at most

$$(2\pi)^{-1}q_n \exp\big(B^2\big)(\lambda_n n^{\frac{1}{2}})^{-1} \int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} \exp\big(-xt\big)h_n^2(x) \exp\big(2(\alpha+n^{-\frac{1}{5}})(\alpha+3x^{\frac{1}{2}})\big)b_n(x)dx$$

$$= 2\pi^{-1}q_n \exp\big(B^2\big)\frac{n^{\frac{3}{2}}}{\lambda_n} \int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} \exp\big(-xt+2(\alpha+n^{-\frac{1}{5}})(\alpha+3x^{\frac{1}{2}})\big)b_n(x)^{-1}dx. \quad (3.20)$$

We now bound the integral appearing in (3.20). In particular, by applying Lemma 23 with $C = 6(\alpha+1), c = 6(\alpha+n^{-\frac{1}{5}})$, and combining with the fact that for all sufficiently large n, $\exp\big((2\alpha+3B)n^{-\frac{1}{5}}\big) \leq 1+n^{-\frac{1}{6}}$ and $\exp\big((36\alpha n^{-\frac{1}{5}}+18n^{-\frac{2}{5}})t^{-1}\big) \leq 1+n^{-\frac{1}{6}}$, we find that for all sufficiently large n, and all $t \geq 1$, the integral appearing in (3.20) is at most

$$(1+n^{-\frac{1}{6}})^3(\frac{\pi}{t\lambda_n})^{\frac{1}{2}} \exp\Big(2\alpha^2+3B\alpha+18\alpha^2t^{-1}-\frac{B^2}{4}t\Big). \tag{3.21}$$

Combining (3.20) and (3.21), we find that the l.h.s. of (3.1) is at most

$$\begin{aligned}
&2\pi^{-1}q_n \exp\big(B^2\big)\frac{n^{\frac{3}{2}}}{\lambda_n}(1+n^{-\frac{1}{6}})^3(\frac{\pi}{t\lambda_n})^{\frac{1}{2}} \exp\Big(2\alpha^2+3B\alpha+18\alpha^2t^{-1}-\frac{B^2}{4}t\Big)\\
\leq\;& 2(\pi t)^{-\frac{1}{2}}(1+n^{-\frac{1}{6}})^4(\frac{n}{\lambda_n})^{\frac{3}{2}} \exp\Big(B^2+2\alpha^2+3B\alpha+18\alpha^2t^{-1}-\frac{B^2}{4}t\Big)\\
\leq\;& 2(\pi t)^{-\frac{1}{2}}(1+n^{-\frac{1}{6}})^4(\frac{n}{\lambda_n})^{\frac{3}{2}} \exp\Big(2(\alpha+B)^2+18\alpha^2t^{-1}-\frac{B^2}{4}t\Big),
\end{aligned}$$

and the first part of Theorem 9 follows, since for $B < B^*$, $t \geq 1$, and all sufficiently large n, $2\pi^{-\frac{1}{2}}(1+n^{-\frac{1}{6}})^4(\frac{n}{\lambda_n})^{\frac{3}{2}} \exp\big(2(\alpha+B)^2+18\alpha^2t^{-1}\big) \leq \exp\big(120(\max(|a_1|,|a_2|)^2+1)\big)$. Similarly, it follows from Lemma 22, Corollary 1, Lemma 24, and Lemma 18 that for all sufficiently large n and all $t \geq 1$, the l.h.s. of (3.2) is at most

$$(2\pi)^{-1}q_n \exp\big(B^2\big)n^{-1}\int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} h_n(x)^2 r_n(x)^{\frac{1}{2}} \exp\big(-xt+2(\alpha+n^{-\frac{1}{5}})(\alpha+3x^{\frac{1}{2}})\big)x^{-1}b_n(x)dx,$$

which is itself at most

$$\pi^{-1}q_n^{\frac{3}{2}} \exp\big(B^2\big)n^{\frac{1}{2}}\int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} \exp\big(-xt+2(\alpha+n^{-\frac{1}{5}})(\alpha+3x^{\frac{1}{2}})\big)x^{-\frac{1}{2}}b_n(x)^{-1}dx. \tag{3.22}$$

It follows from Lemma 3 that $x \geq (n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2$ implies $x^{-\frac{1}{2}} \leq 2B^{-1}$, and thus applying

(3.21), we find that (3.22) is at most

$$2\pi^{-1}q_n^{\frac{3}{2}}\exp\big(B^2\big)n^{\frac{1}{2}}(2B^{-1})\int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2}\exp\big(-xt+2(\alpha+n^{-\frac{1}{5}})(\alpha+3x^{\frac{1}{2}})\big)b_n(x)^{-1}dx$$
$$\leq\ 4(B^2\pi t)^{-\frac{1}{2}}(1+n^{-\frac{1}{6}})^{4.5}(\frac{n}{\lambda_n})^{\frac{1}{2}}\exp\Big(B^2+2\alpha^2+3B\alpha+18\alpha^2t^{-1}-\frac{B^2}{4}t\Big).$$

The second part of Theorem 9 follows, since for all sufficiently large n and all $t \geq 1$, $4(1+n^{-\frac{1}{6}})^{4.5}\pi^{-\frac{1}{2}}(\frac{n}{\lambda_n})^{\frac{1}{2}}\exp\big(B^2+2\alpha^2+3B\alpha+18\alpha^2t^{-1}\big) \leq \exp\big(120(\max(|a_1|,|a_2|)^2+1)\big)$. □

3.4 Comparison to Other Bounds From the Literature

In this subsection we compare the bounds from Theorem 9 to two other explicit bounds given in the literature [119],[23]. In both cases we will prove that the bounds from the literature (applied to $|P^n_{n,\leq n}(t) - P^n_{\leq n}(\infty)|$ for $0 < B < B^*$) grow with n, rendering them impractical in the H-W regime. We begin with the bounds given in [119], which translate to the statement that for each $B \in (0, B^*)$ there exists N_B s.t. $n \geq N_B$ implies that for all $t \geq 0$,
$|P^n_{n,\leq n}(t) - P^n_{\leq n}(\infty)|$ is at most

$$4(n-1)\Big(\sum_{i=1}^{\infty}((\frac{n}{n-1})^i-1)P^n_i(\infty)+((\frac{n}{n-1})^n-1)(1-2P^n_n(\infty))\Big)e^{-\frac{Bn^{\frac{1}{2}}-1}{n-1}t}. \quad (3.23)$$

Note that since $\lim_{n\to\infty}\frac{Bn^{\frac{1}{2}}-1}{n-1} = 0$, the exponential rate of convergence demonstrated by (3.23) goes to zero as $n \to \infty$, rendering the bound in [119] ineffective. We now examine the bounds given in [23], which translate to the statement that for each

$B \in (0, B^*)$ there exists N_B s.t. $n \geq N_B$ implies that for all $t \geq 0$,

$$|P^n_{n,\leq n}(t) - P^n_{\leq n}(\infty)| \leq \left(P^n_n(\infty)^{-1} - 1\right)^{\frac{1}{2}} e^{-\gamma_n t}. \tag{3.24}$$

It is well-known (see [52]) that $\liminf_{n\to\infty} \left(P^n_n(\infty)^{-1} - 1\right)^{\frac{1}{2}} n^{-\frac{1}{4}} > 0$. It follows that the prefactor demonstrated by (3.24) diverges as $n \to \infty$, rendering the bound in [23] ineffective.

It should be noted that although the bounds given in [119] and [23] are ineffective for the particular events of interest in this paper, both bounds hold in much greater generality, and thus remain interesting and applicable in a variety of other settings.

3.5 Conclusion and Open Questions

In this chapter we derived explicit bounds on the distance to steady-state for the $M/M/n$ queue in the H-W regime for the case $B < B^*$, e.g. characterizing the error in estimating the transient probability of delay by the corresponding steady-state quantity. Our bounds hold for any sufficiently large fixed n, scale independently of n in the H-W regime, and (in light of the results of Chapter 2) are tight up to exponential order. We also used our bounds to provide a heuristic rule-of-thumb which a call center manager could use to determine the time it takes an overloaded (underloaded) queueing system to return (probabilistically) to the steady-state. We also studied several known explicit bounds appearing in the literature, and showed that they did not scale favorably with n in the H-W regime.

This work leaves several interesting directions for future research. There are many open questions related to the interaction between weak convergence and convergence to stationarity. In particular, it is an open challenge to derive uniform bounds on the distance to steady-state in the H-W regime for the $M/M/n$ queue with $B > B^*$, and more generally for the $GI/GI/n$ queue with non-Markovian processing times.

3.6 Appendix

3.6.1 Proof of Lemma 23

In this subsection we complete the proof of Lemma 23.

Proof of Lemma 23. By Lemma 3, $(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 < 2\sqrt{\lambda_n n} < (n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2$ for all sufficiently large n. Thus it follows from (3.4) that for all sufficiently large n, the l.h.s. of (3.19) equals

$$\int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{2\sqrt{\lambda_n n}} \exp\big(-xt + cx^{\frac{1}{2}}\big)\big(x - (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)^{-\frac{1}{2}}\big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - x\big)^{-\frac{1}{2}} dx \quad (3.25)$$

$$+ \int_{2\sqrt{\lambda_n n}}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2} \exp\big(-xt + cx^{\frac{1}{2}}\big)\big(x - (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)^{-\frac{1}{2}}\big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - x\big)^{-\frac{1}{2}} dx. (3.26)$$

Let $u_n \triangleq 2\sqrt{\lambda_n n} - (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2$. Since $\big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - x\big)^{-\frac{1}{2}} \leq \big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - 2\sqrt{\lambda_n n}\big)^{-\frac{1}{2}}$ for $x \in \big((n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2, 2\sqrt{\lambda_n n}\big)$, (3.25) is at most

$$\begin{aligned}
&\big((n^{\frac{1}{2}} + \lambda_n^{\frac{1}{2}})^2 - 2\sqrt{\lambda_n n}\big)^{-\frac{1}{2}} \int_{(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2}^{2\sqrt{\lambda_n n}} \exp\big(-xt + cx^{\frac{1}{2}}\big)\big(x - (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)^{-\frac{1}{2}} dx \\
= \;& (\lambda_n + n)^{-\frac{1}{2}} \int_0^{u_n} \exp\Big(-\big(y + (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)t + c\big(y + (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)^{\frac{1}{2}}\Big) y^{-\frac{1}{2}} dy \\
\leq \;& (\lambda_n + n)^{-\frac{1}{2}} \exp\big(-(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 t + c(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})\big) \int_0^{u_n} \exp\big(-yt + cy^{\frac{1}{2}}\big) y^{-\frac{1}{2}} dy,
\end{aligned}$$

since $\big(y + (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2\big)^{\frac{1}{2}} \leq y^{\frac{1}{2}} + n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}}$. It follows from elementary calculus that for all $y > 0$, $-yt + cy^{\frac{1}{2}} \leq -\frac{1}{2}yt + c^2(2t)^{-1}$. Plugging into the above, we find that (3.25) is at most

$$\begin{aligned}
&\big(\lambda_n+n\big)^{-\frac{1}{2}}\exp\big(-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2t+c(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})+c^2(2t)^{-1}\big)\int_0^{u_n}\exp\big(-\frac{1}{2}yt\big)y^{-\frac{1}{2}}dy\\
\leq\ &\big(\lambda_n+n\big)^{-\frac{1}{2}}\exp\big(-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2t+c(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})+c^2(2t)^{-1}\big)\int_0^{\infty}\exp\big(-\frac{1}{2}yt\big)y^{-\frac{1}{2}}dy\\
=\ &\big(\lambda_n+n\big)^{-\frac{1}{2}}\exp\big(-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2t+c(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})+c^2(2t)^{-1}\big)(\frac{2\pi}{t})^{\frac{1}{2}},
\end{aligned}$$

which is itself at most

$$(\frac{\pi}{\lambda_n t})^{\frac{1}{2}}\exp\big(-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2t+c(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})+c^2(2t)^{-1}\big). \qquad (3.27)$$

We now bound (3.26). Since $\big(x-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2\big)^{-\frac{1}{2}} \leq \big(2\sqrt{\lambda_n n}-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2\big)^{-\frac{1}{2}}$ for $x\in\big(2\sqrt{\lambda_n n},(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2\big)$, (3.26) is at most

$$\frac{\sup_{z\in[2\sqrt{\lambda_n n},(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2]}\exp\big(-zt+cz^{\frac{1}{2}}\big)}{\big(2\sqrt{\lambda_n n}-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2\big)^{\frac{1}{2}}}\int_{2\sqrt{\lambda_n n}}^{(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2}\big((n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2-x\big)^{-\frac{1}{2}}dx$$

$$\begin{aligned}
=\ &\frac{\sup_{z\in[2\sqrt{\lambda_n n},(n^{\frac{1}{2}}+\lambda_n^{\frac{1}{2}})^2]}\exp\big(-zt+cz^{\frac{1}{2}}\big)}{\big(2\sqrt{\lambda_n n}-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2\big)^{\frac{1}{2}}}\int_0^{\lambda_n+n}y^{-\frac{1}{2}}dy\\
\leq\ &2(\frac{\lambda_n+n}{3\lambda_n-n})^{\frac{1}{2}}\exp\big(c^2(2t)^{-1}-(\lambda_n n)^{\frac{1}{2}}t\big), \qquad (3.28)
\end{aligned}$$

since $-zt+cz^{\frac{1}{2}}\leq-\frac{1}{2}zt+c^2(2t)^{-1}$, $2\sqrt{\lambda_n n}-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2\geq 3\lambda_n-n$, and $\int_0^{\lambda_n+n}y^{-\frac{1}{2}}dy=2(\lambda_n+n)^{\frac{1}{2}}$. Also, it follows from simple asymptotics that for all sufficiently large n and all $t\geq 1$,

$$2(\frac{\lambda_n+n}{3\lambda_n-n})^{\frac{1}{2}}\exp\big(-(\lambda_n n)^{\frac{1}{2}}t\big)\leq n^{-1}(\frac{\pi}{\lambda_n t})^{\frac{1}{2}}\exp\big(-(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})^2t+c(n^{\frac{1}{2}}-\lambda_n^{\frac{1}{2}})\big). \quad (3.29)$$

Combining with (3.28), it follows that for all sufficiently large n and all $t \geq 1$, (3.26) is at most

$$n^{-1}(\frac{\pi}{\lambda_n t})^{\frac{1}{2}} \exp\big(- (n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}})^2 t + c(n^{\frac{1}{2}} - \lambda_n^{\frac{1}{2}}) + c^2(2t)^{-1}\big). \qquad (3.30)$$

The lemma then follows by using (3.27) to bound (3.25), (3.30) to bound (3.26), applying Lemma 3 and a simple Taylor series expansion, and noting that in all cases 'for all sufficiently large n' can be defined in terms of B and C only (as opposed to c). □

Chapter 4

Asymptotic Scaling and Large Deviations for the Steady-state GI/GI/n Queue in the Halfin-Whitt Regime

4.1 Introduction and Literature Review

Recall from Section 1.3 that the H-W regime was formally introduced by Halfin and Whitt [52], who studied the $GI/M/n$ system (for large n) when the traffic intensity scales like $1 - Bn^{-\frac{1}{2}}$ for some strictly positive B. They proved that under minor technical assumptions on the inter-arrival distribution, this sequence of $GI/M/n$ queueing models has the following properties:

(i) the steady-state probability that an arriving job finds all servers busy (i.e. the probability of wait) has a non-trivial limit;

(ii) the sequence of queueing processes, normalized by $n^{\frac{1}{2}}$, converges weakly to a non-trivial positive recurrent diffusion;

(iii) the sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight and converges distributionally to the mixture of a point mass at 0 and an exponential distribution.

Similar weak convergence results under the H-W scaling were subsequently obtained for more general multi-server systems [87], [60], [75], [43], [91] with the most general results appearing in [91] (and follow-up papers [90],[88]). As the theory of weak convergence generally relies heavily on the assumption of compact time intervals, the most general of these results hold only in the transient regime. Indeed, with the exception of [52] (which treats exponential processing times), [60] (which treats deterministic processing times), and [43] (which treats processing times with finite support), all of the aforementioned results are for the associated sequence of normalized *transient* queue length distributions only, leaving many open questions about the associated *steady-state* queue length distributions.

In particular, in [43] it is shown for the case of processing times with finite support that the sequence of steady-state queue length distributions (normalized by $n^{\frac{1}{2}}$) is tight, and has a limit whose tail decays exponentially fast. The authors further prove that this exponential rate of decay (i.e. large deviation exponent) is $-2B(c_A^2 + c_S^2)^{-1}$, where B is the spare capacity parameter, and c_A^2, c_S^2 are the squared coefficients of variation of the inter-arrival and processing time distributions. In [43] it was conjectured that this result should hold for more general processing time distributions. However, prior to this thesis no further progress on this question has been achieved.

In this chapter we resolve the conjectures made in [43] w.r.t. tightness of the steady-state queue length, and take a large step towards resolving the conjectures made w.r.t. the large deviation exponent. We prove that as long as the inter-arrival and processing time distributions satisfy very minor technical conditions (e.g. finite

$2+\epsilon$ moments), the associated sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight. Under the same minor technical conditions we derive an upper bound on the large deviation exponent of the limiting steady-state queue length matching that conjectured by Gamarnik and Momcilovic in [43]. We also prove a matching lower bound when the arrival process is Poisson.

Our main proof technique is the derivation of new and simple bounds for the FCFS $GI/GI/n$ queue. Our bounds are of a structural nature, hold for all n and all times $t \geq 0$, and have intuitive closed-form representations as the suprema of certain natural processes which converge weakly to Gaussian processes. Our upper and lower bounds also exhibit a certain duality relationship, and exemplify a general methodology which may be useful for analyzing a variety of queueing systems. We further illustrate the utility of this methodology by deriving the first non-trivial bounds for the weak-limit process studied in [91].

We note that our techniques allow us to analyze many properties of the $GI/GI/n$ queue in the H-W regime without having to consider the complicated exact dynamics of the $GI/GI/n$ queue. Interestingly, such ideas were used in the original paper of Halfin and Whitt [52] to show tightness of the steady-state queue length for the $GI/M/n$ queue under the H-W scaling, but do not seem to have been used in subsequent works on queues in the H-W regime.

4.1.1 Outline of chapter

The rest of the chapter proceeds as follows. In Section 4.2, we present our main results. In Section 4.3, we establish our general-purpose upper bounds for the queue length in a properly initialized FCFS $GI/GI/n$ queue. In Section 4.4, we establish our general-purpose lower bounds for the queue length in a properly initialized FCFS $M/GI/n$ queue. In Section 4.5 we use our bounds to prove the tightness of the steady-state queue length when the system is in the H-W regime. In Section 4.6 we combine our bounds with known results about weak limits and the suprema of Gaussian processes to prove our large deviation results. In Section 4.7 we use our bounds to study the weak limit derived in [91]. In Section 4.8 we summarize our main results and comment on directions for future research. We include a technical appendix in Section 4.9.

4.2 Main Results

Recall from Chapter 1 that $\lambda_n = n - Bn^{\frac{1}{2}}$, and Q^n is the First-Come-First-Serve (FCFS) $GI/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as $A\lambda_n^{-1}$ and processing times drawn i.i.d. distributed as S (initial conditions will be specified later). Suppose that $\mathbb{E}[A] = \mu_A^{-1} < \infty, \mathbb{E}[S] = \mu_S^{-1} < \infty$, and $\mathbb{P}(A = 0) = \mathbb{P}(S = 0) = 0$ Recall that σ_A^2 and σ_S^2 denote the variances of A and S, respectively, and c_A^2 and c_S^2 denote the squared coefficients of variation (s.c.v.) of A and S, respectively. All processes should be assumed right-continuous with left limits (r.c.l.l.) unless stated otherwise. All empty summations should be evaluated as zero, and all empty products should be evaluated as one.

4.2.1 Main results

Our main results will require two additional sets of assumptions on A and S. The first set of assumptions, which we call the H-W assumptions, ensures that $\{Q^n(t), n \geq 1\}$ is in the H-W scaling regime as $n \rightarrow \infty$. We say that A and S satisfy the H-W assumptions iff $\mu_A = \mu_S$, in which case we denote this common rate by μ. The second set of assumptions, which we call the T_0 assumptions, is a set of additional minor technical conditions we require for our main results.

(i) There exists $\epsilon > 0$ s.t. $\mathbb{E}[A^{2+\epsilon}], \mathbb{E}[S^{2+\epsilon}] < \infty$.

(ii) $c_A^2 + c_S^2 > 0$. Namely either A or S is a non-trivial r.v.

(iii) $\limsup_{t \downarrow 0} t^{-1}\mathbb{P}(S \leq t) < \infty$.

(iv) For all sufficiently large n, $Q^n(t)$ converges weakly to a steady-state distribution $Q^n(\infty)$ as $t \to \infty$.

We note that technical condition (iii) of the T_0 assumptions is not very restrictive, and is (for example) satisfied by any discrete distribution with no mass at zero, and any continuous distribution with finite density at zero. Furthermore, condition (iii) is in some sense natural, as certain closely related tightness results from the literature are known to require a similar condition (see the discussion in [110]). We refer the interested reader to [4] for an excellent discussion of technical condition (iv), which is also not very restrictive.

We now state our main results. We begin by establishing the tightness of the steady-state queue length for the FCFS $GI/GI/n$ queue in the H-W regime.

Theorem 11. *If A and S satisfy the H-W and T_0 assumptions, then the sequence $\{(Q^n(\infty) - n)^+ n^{-\frac{1}{2}}, n \geq 1\}$ is tight.*

In words, the queue length $(Q^n(\infty) - n)^+$ scales like $O(n^{\frac{1}{2}})$.

We now establish an upper bound for the large deviation exponent of the limiting steady-state queue length for the FCFS $GI/GI/n$ queue in the H-W regime, and a matching lower bound when the arrival process is Poisson.

Theorem 12. *Under the same assumptions as Theorem 11,*

$$\limsup_{x\to\infty} x^{-1}\log\left(\limsup_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)^+ n^{-\frac{1}{2}} > x\Big)\right) \le -2B(c_A^2+c_S^2)^{-1}.$$

If in addition A is an exponentially distributed r.v., namely the system is $M/GI/n$, then

$$\begin{aligned}&\lim_{x\to\infty} x^{-1}\log\left(\liminf_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)^+ n^{-\frac{1}{2}} > x\Big)\right)\\ = &\lim_{x\to\infty} x^{-1}\log\left(\limsup_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)^+ n^{-\frac{1}{2}} > x\Big)\right) \quad = \quad -2B(c_A^2+c_S^2)^{-1}.\end{aligned}$$

In words, Theorem 12 states that the tail of the limiting steady-state queue length is bounded from above by $\exp\big(-2B(c_A^2+c_S^2)^{-1}x+o(x)\big)$; and when the arrival process is Poisson, the tail of the limiting steady-state queue length is bounded from below by $\exp\big(-2B(c_A^2+c_S^2)^{-1}x-o(x)\big)$, where $o(x)$ is some non-negative function s.t. $\lim_{x\to\infty} x^{-1}o(x) = 0$. Note that Theorem 12 translates into bounds for the large deviation behavior of any weak limit of the sequence $\{\big(\mathrm{Q}^n(\infty)-n\big)^+ n^{-\frac{1}{2}}, n \ge 1\}$, where at least one weak limit exists by Theorem 11.

4.3 Upper Bound

In this section, we prove a general upper bound for the FCFS $GI/GI/n$ queue, when properly initialized. The bound is valid for all finite n, and works in both the transient and steady-state (when it exists) regimes. Although we will later customize this bound to the H-W regime to prove our main results, we note that the bound is in no way limited to that regime. For a non-negative r.v. X with finite mean $\mathbb{E}[X] > 0$, let $R(X)$ denote a r.v. distributed as the residual life distribution of X. Namely, for all $z \geq 0$,

$$\mathbb{P}\big(R(X) > z\big) = (\mathbb{E}[X])^{-1} \int_{z}^{\infty} \mathbb{P}(X > y)dy. \tag{4.1}$$

Recall that associated with a r.v. X, an equilibrium renewal process with renewal distribution X is a counting process in which the first inter-event time is distributed as $R(X)$, and all subsequent inter-event times are drawn i.i.d. distributed as X; an ordinary renewal process with renewal distribution X is a counting process in which all inter-event times, including the first, are drawn i.i.d. distributed as X. Let $\{N_i(t), i = 1, \ldots, n\}$ denote a set of n i.i.d. equilibrium renewal processes with renewal distribution S. Let $A(t)$ denote an independent equilibrium renewal process with renewal distribution A.

Let $\mathcal{Q}$ denote the FCFS $GI/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as A, processing times drawn i.i.d. distributed as S, and the following initial conditions. For $i = 1, \ldots, n$, there is a single job initially being processed on server i, and the set of initial processing times of these n initial jobs is drawn i.i.d. distributed as $R(S)$. There are zero jobs waiting in queue, and the first inter-arrival time is distributed as $R(A)$, independent of the initial processing times of those jobs initially in system. We now establish an upper bound for $Q(t)$, the number in system at time t in $\mathcal{Q}$.

Theorem 13. *For all $x > 0$, and $t \geq 0$,*

$$\mathbb{P}\big((Q(t) - n)^+ > x\big) \leq \mathbb{P}\Big(\sup_{0 \leq s \leq t} \big(A(s) - \sum_{i=1}^{n} N_i(s)\big) > x\Big).$$

If in addition $Q(t)$ converges weakly to a steady-state distribution $Q(\infty)$ as $t \to \infty$, then for all $x > 0$,

$$\mathbb{P}\big((Q(\infty) - n)^+ > x\big) \leq \mathbb{P}\Big(\sup_{t \geq 0} \big(A(t) - \sum_{i=1}^{n} N_i(t)\big) > x\Big).$$

Note that our bounds are monotone in time, as when t increases the supremum appearing in Theorem 13 is taken over a larger time window, and the bound for the steady-state is the natural limit of these transient bounds.

We will prove Theorem 13 by analyzing a different FCFS $G/GI/n$ queue $\tilde{Q}$ which represents a 'modified' FCFS $GI/GI/n$ queue, in which all servers are kept busy at all times by adding artificial arrivals whenever a server would otherwise go idle. We note that our construction is similar to several constructions appearing in the literature. Our bounding system is closely related to the so-called Queue with Autonomous Service, a model studied previously by several authors [9],[57],[113],[66]. Another related work is [20], in which the queue length of the $G/GI/1$ queue is bounded by considering a modified system in which the server goes on a vacation whenever it would have otherwise gone idle. Also, in [52], the queue length of the $GI/M/n$ queue is bounded by considering a modified system in which a reflecting barrier is placed at state n.

We now construct the FCFS $G/GI/n$ queue $\tilde{Q}$ on the same probability space as $\{N_i(t), i = 1, \ldots, n\}$ and $A(t)$. We begin by defining two auxiliary processes $\tilde{A}(t)$ and $\tilde{Q}(t)$, where $\tilde{A}(t)$ will become the arrival process to $\tilde{Q}$, and we will later prove that $\tilde{Q}(t)$ equals the number in system in $\tilde{Q}$ at time t. Let $\tau_0 \triangleq 0$, $\{\tau_k, k \geq 1\}$ denote the sequence of event times in the pooled renewal process $A(t) + \sum_{i=1}^{n} N_i(t)$, $dA(t) \triangleq A(t) - A(t^-)$, $A(s,t) \triangleq A(t) - A(s)$, and $dN_i(t) \triangleq N_i(t) - N_i(t^-)$, $N_i(s,t) \triangleq N_i(t) - N_i(s)$ for $i = 1, \ldots, n$.

We now define the processes $\tilde{A}(t)$ and $\tilde{Q}(t)$ inductively over $\{\tau_k, k \geq 0\}$. Let $\tilde{A}(\tau_0) \triangleq 0$, $\tilde{Q}(\tau_0) \triangleq n$. Now suppose that for some $k \geq 0$, we have defined $\tilde{A}(t)$ and $\tilde{Q}(t)$ for all $t \leq \tau_k$. We now define these processes for $t \in (\tau_k, \tau_{k+1}]$. For $t \in (\tau_k, \tau_{k+1})$, let $\tilde{A}(t) \triangleq \tilde{A}(\tau_k)$, and $\tilde{Q}(t) \triangleq \tilde{Q}(\tau_k)$. Note that w.p.1 $dA(\tau_{k+1}) + \sum_{i=1}^{n} dN_i(\tau_{k+1}) = 1$, since $R(A)$ and $R(S)$ are continuous r.v.s, $\mathbb{P}(A = 0) = \mathbb{P}(S = 0) = 0$, and

$A(t), \{N_i(t), i = 1, \ldots, n\}$ are mutually independent. We define

$$\tilde{A}(\tau_{k+1}) \triangleq \begin{cases} \tilde{A}(\tau_k) + 1 & \text{if } dA(\tau_{k+1}) = 1; \\ \tilde{A}(\tau_k) + 1 & \text{if } \sum_{i=1}^n dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) \leq n; \\ \tilde{A}(\tau_k) & \text{otherwise (i.e. } \sum_{i=1}^n dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) > n). \end{cases}$$

Similarly, we define

$$\tilde{Q}(\tau_{k+1}) \triangleq \begin{cases} \tilde{Q}(\tau_k) + 1 & \text{if } dA(\tau_{k+1}) = 1; \\ \tilde{Q}(\tau_k) & \text{if } \sum_{i=1}^n dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) \leq n; \\ \tilde{Q}(\tau_k) - 1 & \text{otherwise (i.e. } \sum_{i=1}^n dN_i(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) > n). \end{cases}$$

Combining the above completes our inductive definition of $\tilde{A}(t)$ and $\tilde{Q}(t)$. Since w.p.1 $\lim_{k \to \infty} \tau_k = \infty$, it follows that w.p.1 both $\tilde{A}(t)$ and $\tilde{Q}(t)$ are well-defined on $[0, \infty)$. We note that it also follows from our construction that w.p.1 both $\tilde{A}(t)$ and $\tilde{Q}(t)$ are r.c.l.l., and define $d\tilde{A}(t) \triangleq \tilde{A}(t) - \tilde{A}(t^-)$. It also follows from our construction that the dynamics of $\tilde{Q}$ are identical to those of the so-called Queue with Autonomous Service, a model studied previously by several authors [9],[57],[113],[66]; we refer the reader to [113] for details.

We now construct the FCFS $G/GI/n$ queue $\tilde{\mathcal{Q}}$ using the auxiliary process $\tilde{A}(t)$. Let V_i^j denote the length of the jth renewal interval in process $N_i(t), j \geq 1, i = 1, \ldots, n$. Then $\tilde{\mathcal{Q}}$ is defined to be the FCFS $G/GI/n$ queue with arrival process $\tilde{A}(t)$ and processing time distribution S, where the jth job assigned to server i (after time 0) is assigned processing time V_i^{j+1} for $j \geq 1, i = 1, \ldots, n$. The initial conditions for $\tilde{\mathcal{Q}}$ are s.t. for $i = 1, \ldots, n$, there is a single job initially being processed on server i with initial processing time V_i^1, and there are zero jobs waiting in queue.

We now analyze $\tilde{\mathcal{Q}}$, proving that

Lemma 25. *For $i = 1, \ldots, n$, exactly one job departs from server i at each time $t \in \{\sum_{l=1}^{j} V_i^l, j \geq 1\}$, and there are no other departures from server i. Also, no server ever idles in $\tilde{Q}$, $\tilde{Q}(t)$ equals the number in system in $\tilde{Q}$ at time t for all $t \geq 0$, and for all $k \geq 1$,*

$$\tilde{Q}(\tau_k) - n = \max\left(0, \tilde{Q}(\tau_{k-1}) - n + dA(\tau_k) - \sum_{i=1}^{n} dN_i(\tau_k)\right). \tag{4.2}$$

Proof. The proof proceeds by induction on $\{\tau_k, k \geq 0\}$, with induction hypothesis that the lemma holds for all $t \leq \tau_k$. The base case $k = 0$ follows from the the initial conditions of $\tilde{Q}$ and $\tilde{Q}(t)$. Thus assume that the induction hypothesis holds for some fixed $k \geq 0$. We first establish the induction step for the statements about the departure process and non-idling of servers. Let us fix some $i \in \{1, \ldots, n\}$. By the induction hypothesis, server i was non-idling on $[0, \tau_k]$, and the set of departure times from server i on $[0, \tau_k]$ was exactly $\{\sum_{l=1}^{j} V_i^l, j = 1, \ldots, N_i(\tau_k)\}$. We claim that the next departure from server i occurs at time $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l$. Indeed, if $N_i(\tau_k) = 0$, the next departure from server i is the first departure from server i, which occurs at time V_i^1. If instead $N_i(\tau_k) > 0$, then the last departure from server i to occur at or before time τ_k occured at time $\sum_{l=1}^{N_i(\tau_k)} V_i^l$. At that time a new job began processing on server i with processing time $V_i^{N_i(\tau_k)+1}$. This job will depart at time $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l$, verifying the claim. It follows that no server idles on (τ_k, τ_{k+1}), since $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l \in \{\tau_j, j \geq 1\}$, and thus $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l \geq \tau_{k+1}$. We now treat two cases. First, suppose $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l > \tau_{k+1}$. Then there are no departures from server i on $(\tau_k, \tau_{k+1}]$ and the induction step follows immediately from the induction hypothesis. Alternatively, suppose $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l = \tau_{k+1}$. In this case the next departure from server i occurs at time τ_{k+1}, $dN_i(\tau_{k+1}) = 1$, and all other servers are non-idling and have no departures on $(\tau_k, \tau_{k+1}]$. Thus if there are at least $n + 1$ jobs in $\tilde{Q}$ at time τ_k, then there are at least $n + 1$ jobs in $\tilde{Q}$ at time τ_{k+1}^-, and some job begins processing on server i at time τ_{k+1}. Alternatively, if there are exactly n jobs in $\tilde{\mathcal{Q}}$ at time τ_k, then $\tilde{Q}(\tau_k) = n$ by the induction hypothesis. Thus $d\tilde{A}(\tau_{k+1}) = 1$, and this arrival immediately begins processing on server i. Combining the above treats all cases since there are at least n jobs in $\tilde{\mathcal{Q}}$ at time τ_k by the induction hypothesis, completing the induction step.

We now prove the induction step for the statement that $\tilde{Q}(t)$ equals the number in system in $\tilde{\mathcal{Q}}$ at time t, as well as (4.2). Since we have already proven that any departures from $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$ occur at time τ_{k+1}, and by construction any jumps in $\tilde{A}(t)$ and $\tilde{Q}(t)$ on $(\tau_k, \tau_{k+1}]$ occur at time τ_{k+1}, it suffices to prove that $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1}. First, suppose $dA(\tau_{k+1}) = 1$. Then $\sum_{i=1}^n dN_i(\tau_{k+1}) = 0$, $\tilde{Q}(\tau_k) \geq n$ by the induction hypothesis, and $\tilde{Q}(\tau_{k+1}) = \tilde{Q}(\tau_k)+1$. Thus

$$\begin{aligned}\max\left(0, \tilde{Q}(\tau_k) - n + dA(\tau_{k+1}) - \textstyle\sum_{i=1}^n dN_i(\tau_{k+1})\right) &= \max\left(0, \tilde{Q}(\tau_k) - n + 1\right) \\ &= \tilde{Q}(\tau_k) - n + 1 = \tilde{Q}(\tau_{k+1}) - n,\end{aligned}$$

showing that (4.2) holds. Note that $\sum_{i=1}^n dN_i(\tau_{k+1}) = 0$ implies that $\sum_{l=1}^{N_i(\tau_k)+1} V_i^l > \tau_{k+1}$ for all $i = 1, \ldots, n$, and we have already proven that in this case there are no departures from $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$. Since $dA(\tau_{k+1}) = 1$ implies $d\tilde{A}(\tau_{k+1}) = 1$, it follows that the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} is one more than the number in system in $\tilde{\mathcal{Q}}$ at time τ_k. Thus $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} by the induction hypothesis.

Now suppose that $\sum_{i=1}^n dN_i(\tau_{k+1}) = 1$. Then $dA(\tau_{k+1}) = 0$, and there exists a unique index i^* s.t. $\sum_{l=1}^{N_{i^*}(\tau_k)+1} V_{i^*}^l = \tau_{k+1}$. We have already proven that in this case there are no departures from $\tilde{\mathcal{Q}}$ on (τ_k, τ_{k+1}), and a single departure from $\tilde{\mathcal{Q}}$ at time τ_{k+1} (on server i^*). First suppose that there are at least $n+1$ jobs in $\tilde{\mathcal{Q}}$ at time τ_k.

Then $\tilde{Q}(\tau_k) \geq n+1$ by the induction hypothesis, and $\tilde{Q}(\tau_{k+1}) = \tilde{Q}(\tau_k) - 1$. Thus

$$\begin{aligned}\max\left(0, \tilde{Q}(\tau_k) - n + dA(\tau_{k+1}) - \textstyle\sum_{i=1}^n dN_i(\tau_{k+1})\right) &= \max\left(0, \tilde{Q}(\tau_k) - n - 1\right) \\ &= \tilde{Q}(\tau_k) - n - 1 = \tilde{Q}(\tau_{k+1}) - n,\end{aligned}$$

showing that (4.2) holds. Since $d\tilde{A}(\tau_{k+1}) = 0$, there are no arrivals to $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$. Combining the above, we find that the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} is one less than the number in system in $\tilde{\mathcal{Q}}$ at time τ_k. Thus $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} by the induction hypothesis.

Alternatively, suppose that $\sum_{i=1}^n dN_i(\tau_{k+1}) = 1$ and there are exactly n jobs in $\tilde{\mathcal{Q}}$ at time τ_k. Then $\tilde{Q}(\tau_k) = n$ by the induction hypothesis, and $\tilde{Q}(\tau_{k+1}) = \tilde{Q}(\tau_k)$. Thus

$$\begin{aligned}\max\left(0, \tilde{Q}(\tau_k) - n + dA(\tau_{k+1}) - \textstyle\sum_{i=1}^n dN_i(\tau_{k+1})\right) &= \max\left(0, \tilde{Q}(\tau_k) - n - 1\right) \\ &= 0 \quad = \quad \tilde{Q}(\tau_{k+1}) - n,\end{aligned}$$

showing that (4.2) holds. Since $d\tilde{A}(\tau_{k+1}) = 1$, there is a single arrival to $\tilde{\mathcal{Q}}$ on $(\tau_k, \tau_{k+1}]$. Combining the above, we find that the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_k. Thus $\tilde{Q}(\tau_{k+1})$ equals the number in system in $\tilde{\mathcal{Q}}$ at time τ_{k+1} by the induction hypothesis. Since $\tilde{Q}(\tau_k) \geq n$ by the induction hypothesis, this treats all cases, completing the proof of the induction and the lemma. □

We now 'unfold' recursion (4.2) to derive a simple one-dimensional random walk representation for $\tilde{Q}(t)$. We note that the relationship between recursions such as (4.2) and the suprema of associated one-dimensional random walks is well-known (see [9],[20]). Then it follows from (4.2) and a straightforward induction on $\{\tau_k, k \geq 0\}$

that w.p.1, for all $k \geq 0$,

$$\tilde{Q}(\tau_k) - n = \max_{0 \leq j \leq k} \Big(A\big(\tau_{k-j}, \tau_k\big) - \sum_{i=1}^{n} N_i\big(\tau_{k-j}, \tau_k\big) \Big).$$

As all jumps in $\tilde{Q}(t)$ occur at times $t \in \{\tau_k, k \geq 1\}$, it follows that

Corollary 3. *W.p.1, for all $t \geq 0$,*

$$\tilde{Q}(t) - n = \sup_{0 \leq s \leq t} \Big(A\big(t-s, t\big) - \sum_{i=1}^{n} N_i\big(t-s, t\big) \Big).$$

We now prove that $\tilde{Q}(t)$ provides an upper bound for $Q(t)$.

Proposition 3. *$Q(t)$ and $\tilde{Q}(t)$ can be constructed on the same probability space so that w.p.1 $Q(t) \leq \tilde{Q}(t)$ for all $t \geq 0$.*

For our later results, it will be useful to first prove a general comparison result for $G/G/n$ queues. Although such results seem to be generally known in the queueing literature (see [109],[97]), we include a proof for completeness. For an event E, let $I(E)$ denote the indicator function of E.

Lemma 26. *Let Q^1 and Q^2 be two FCFS $G/G/n$ queues with finite, strictly positive inter-arrival and processing times. Let $\{T^i_k, k \geq 1\}$ denote the ordered sequence of arrival times to Q^i, $i \in \{1, 2\}$. Let S^i_k denote the processing time assigned to the job that arrives to Q^i at time T^i_k, $k \geq 1, i \in \{1, 2\}$. Further suppose that*

(i) The initial number in system in Q^1 is at most n;

(ii) For each job J initially in Q^1 there is a distinct corresponding job J' initially in Q^2 s.t. the initial processing time of J in Q^1 equals the initial processing time of J' in Q^2;

(iii) $\{T_k^1, k \geq 1\}$ *is a subsequence of* $\{T_k^2, k \geq 1\}$*;*

(iv) For all $k \geq 1$*, the job that arrives to* $\mathcal{Q}^2$ *at time* T_k^1 *is assigned processing time* S_k^1*, the same processing time assigned to the job which arrives to* $\mathcal{Q}^1$ *at that time.*

Then the number in system in $\mathcal{Q}^2$ *at time* t *is at least the number in system in* $\mathcal{Q}^1$ *at time* t *for all* $t \geq 0$*.*

Proof. Let $Z^i(t)$ denote the number of jobs initially in $\mathcal{Q}^i$ which are still in $\mathcal{Q}^i$ at time t, $i \in \{1, 2\}$. We claim that $Z^2(t) \geq Z^1(t)$ for all $t \geq 0$. Indeed, let J be any job initially in $\mathcal{Q}^1$, and let S_J denote its initial processing time. Then (ii) ensures the existence of a distinct corresponding job J' initially in $\mathcal{Q}^2$, with the same initial processing time S_J. Since by (i) all jobs initially in Q^1 begin processing at time 0, it follows that J departs $\mathcal{Q}^1$ at time S_J, while J' departs $\mathcal{Q}^2$ no earlier than S_J. Making this argument for each job J initially in $\mathcal{Q}^1$ proves that $Z^2(t) \geq Z^1(t)$ for all $t \geq 0$.

Let D_k^i denote the time at which the job that arrives to $\mathcal{Q}^i$ at time T_k^1 departs from $\mathcal{Q}^i$, $k \geq 1, i \in \{1, 2\}$. We now prove by induction that for $k \geq 1$, $D_k^2 \geq D_k^1$, from which the proposition follows. Observe that for all $k \geq 1$,

$$D_k^1 = \inf\{t: \ t \geq T_k^1, Z^1(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n - 1\} + S_k^1. \tag{4.3}$$

Also,

$$D_k^2 \geq \inf\{t: \ t \geq T_k^1, Z^1(t) + \sum_{j=1}^{k-1} I(D_j^2 > t) \leq n - 1\} + S_k^1, \tag{4.4}$$

where the inequality in (4.4) arises since $Z^2(t) \geq Z^1(t)$ for all $t \geq 0$, and the job that arrives to $\mathcal{Q}^2$ at time T_k^1 may have to wait for additional jobs, which either were initially present in $\mathcal{Q}^2$ but not $\mathcal{Q}^1$, or which arrive at a time belonging to $\{T_k^2, k \geq 1\} \setminus \{T_k^1, k \geq 1\}$.

For the base case $k = 1$, note that $D_1^1 = \inf\{t : t \geq T_1^1, Z^1(t) \leq n-1\} + S_1^1$, while $D_1^2 \geq \inf\{t : t \geq T_1^1, Z^1(t) \leq n-1\} + S_1^1$.

Now assume the induction is true for all $j \leq k$. Then for all $t \geq 0$, $\sum_{j=1}^k I(D_j^2 > t) \geq \sum_{j=1}^k I(D_j^1 > t)$. Thus

$$\inf\{t : \ t \geq T_{k+1}^1, Z^1(t) + \sum_{j=1}^k I(D_j^1 > t) \leq n-1\} + S_{k+1}^1$$

$$\leq \inf\{t : \ t \geq T_{k+1}^1, Z^1(t) + \sum_{j=1}^k I(D_j^2 > t) \leq n-1\} + S_{k+1}^1.$$

It then follows from (5.3) and (4.4) that $D_{k+1}^1 \leq D_{k+1}^2$, completing the induction. □

We now complete the proof of Proposition 3.

Proof of Proposition 3. We construct $\tilde{Q}$ and Q on the same probability space. We assign Q and $\tilde{Q}$ the same initial conditions, and let $A(t)$ be the arrival process to Q on $(0, \infty)$. Let $\{t_k, k \geq 1\}$ denote the ordered sequence of event times in $A(t)$. It follows from the construction of $\tilde{A}(t)$ that $\{t_k, k \geq 1\}$ is a subsequence of the set of event times in $\tilde{A}(t)$. We let the processing time assigned to the arrival to $\tilde{Q}$ at time t_k equal the processing time assigned to the arrival to Q at time t_k, $k \geq 1$. It follows that w.p.1 Q and $\tilde{Q}$ satisfy the conditions of Lemma 26. Combining the above with Lemma 25 completes the proof. □

We now complete the proof of Theorem 13.

Proof of Theorem 13. By elementary renewal theory (see [27]), $A(s)_{0 \leq s \leq t}$ has the same distribution (on the process level) as $A\big(t-s,t\big)_{0 \leq s \leq t}$, and $\sum_{i=1}^n N_i(s)_{0 \leq s \leq t}$ has the same distribution (on the process level) as $\sum_{i=1}^n N_i\big(t-s,t\big)_{0 \leq s \leq t}$. Combining with the independence of $A(t)$ and $\sum_{i=1}^n N_i(t)$, Corollary 7, and Proposition 3, proves the theorem.

We now prove the corresponding steady-state result. Note that for any $x > 0$, the sequence of events $\big\{\sup_{0\leq s\leq t}\big(A(s) - \sum_{i=1}^{n} N_i(s)\big) > x, t \geq 0\big\}$ is monotonic in t. It follows from the continuity of probability measures that

$$\lim_{t\to\infty} \mathbb{P}\bigg(\sup_{0\leq s\leq t}\big(A(s) - \sum_{i=1}^{n} N_i(s)\big) > x\bigg) = \mathbb{P}\bigg(\sup_{t\geq 0}\big(A(t) - \sum_{i=1}^{n} N_i(t)\big) > x\bigg).$$

The steady-state result then follows from the corresponding transient result and the definition of weak convergence, since $Q(\infty)$ has integer support. □

4.4 Lower Bound

In this section, we prove a general lower bound for the $M/GI/n$ queue, when properly initialized. Suppose A is an exponentially distributed r.v. Let Z denote a Poisson r.v. with mean $\frac{\mu_A}{\mu_S}$. Let Q_2 denote the $M/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as A, processing times drawn i.i.d. distributed as S, and the following initial conditions. At time 0 there are Z jobs in system. This set of initial jobs have initial processing times drawn i.i.d. distributed as $R(S)$, independent of Z. If $Z \geq n$, a set of exactly n initial jobs is selected uniformly at random (u.a.r.) to be processed initially, and the remaining initial jobs queue for processing. Suppose also that the first inter-arrival time is distributed as $R(A)$ (also an exponentially distributed r.v.) independent of both Z and the initial processing times of those jobs initially in the system. Recall the processes $A(t)$ and $\{N_i(t), i = 1, \dots, n\}$, which were defined previously at the start of Section 4.3. Then $Q_2(t)$, the number in system at time t in Q_2, satisfies

Theorem 14. *For all $x > 0$, and $t \geq 0$,*

$$\mathbb{P}\big((Q_2(t) - n)^+ > x\big) \geq \mathbb{P}(Z \geq n) \sup_{0 \leq s \leq t} \mathbb{P}\big(A(s) - \sum_{i=1}^{n} N_i(s) > x\big).$$

If in addition $Q_2(t)$ converges weakly to a steady-state distribution $Q(\infty)$ as $t \to \infty$, then for all $x > 0$,

$$\mathbb{P}\big((Q(\infty) - n)^+ > x\big) \geq \mathbb{P}(Z \geq n) \sup_{t \geq 0} \mathbb{P}\big(A(t) - \sum_{i=1}^{n} N_i(t) > x\big).$$

Comparing with Theorem 13, we see that our upper and lower bounds exhibit a certain duality, marked by the order of the $\mathbb{P}$ and sup operators.

We will prove Theorem 14 by coupling $\mathcal{Q}_2$ to *both* an associated FCFS $M/GI/\infty$ queue $\mathcal{Q}_\infty$ and a certain family of FCFS $G/G/n$ queues $\{\mathcal{Q}_2^s, s \geq 0\}$. For each $s \geq 0$, our coupling ensures that $Q_2^s(t)$, the number in system at time t in $\mathcal{Q}_2^s$, provides a lower bound for $Q_2(t)$ for all $t \geq s$, and that the set of remaining processing times (at time s) of those jobs in $\mathcal{Q}_2^s$ at time s is a random thinning of the set of remaining processing times (at time s) of those jobs in $\mathcal{Q}_\infty$ at time s. We note that some of the ideas involved in the proof of our lower bound have appeared in the literature before (see [99], [97], [112]).

We now construct $\mathcal{Q}_\infty$ and $\{\mathcal{Q}_2^s, s \geq 0\}$. We assign $\mathcal{Q}_\infty$ the same initial conditions as $\mathcal{Q}_2$ (although in $\mathcal{Q}_\infty$ all initial jobs begin processing at time 0). We let $\mathcal{Q}_\infty$ and $\mathcal{Q}_2$ have the same arrival process, and for each arrival, we let the processing time assigned to this arrival to $\mathcal{Q}_\infty$ equal the processing time assigned to this arrival to $\mathcal{Q}_2$.

We now describe the initial conditions and arrival process for $\mathcal{Q}_2^s$ in terms of an appropriate thinning of the initial conditions and arrival process of $\mathcal{Q}_\infty$, where the nature of this thinning depends on $Q_\infty(s)$, the number in system at time s in $\mathcal{Q}_\infty$.

If $Q_\infty(s) < n$, then the initial conditions of Q_2^s are to have zero jobs in system, and the arrival process to Q_2^s is to have zero arrivals on $[0, \infty)$. If $Q_\infty(s) \geq n$, then we select a size-n subset $\mathcal{C}^s$ of jobs u.a.r. from all subsets of the jobs being processed in Q_∞ at time s. Let $\mathcal{C}_0^s$ denote those jobs in $\mathcal{C}^s$ which were initially in Q_∞ at time 0. Then the initial conditions of Q_2^s are as follows. For each job $J \in \mathcal{C}_0^s$, there is a corresponding job J' initially in Q_2^s, where the initial processing time of J' in Q_2^s equals the initial processing time of J in Q_∞. There are no other initial jobs in Q_2^s. The arrival process to Q_2^s on $(0, s]$ is as follows. For each job J that arrives to Q_∞ (and thus to Q_2) on $(0, s]$, say at time τ, there is a corresponding arrival J' to Q_2^s at time τ iff $J \in \mathcal{C}^s \setminus \mathcal{C}_0^s$. In this case, the processing time assigned to J' in Q_2^s equals the processing time assigned to J in Q_∞. There are no other arrivals to Q_2^s on $(0, s]$. We let Q_2^s, Q_∞, and Q_2 have the same arrival process on (s, ∞), and for each arrival, we let the processing time assigned to this arrival to Q_2^s equal the processing time assigned to this arrival to Q_∞(and thus Q_2).

We claim that our coupling of Q_∞ to Q_2 and construction of Q_2^s ensure that Q_2^s and Q_2 satisfy the conditions of Lemma 26. Indeed, for each job initially in Q_2^s, there is a distinct corresponding job initially in Q_2 with the same initial processing time. Also, for each job that arrives to Q_2^s, there is a distinct corresponding job that arrives to Q_2 at the same time with the same processing time. Thus w.p.1 $Q_2^s(t)$, the number in system at time t in Q_2^s, satisfies

$$Q_2(t) \geq Q_2^s(t) \text{ for all } s, t \geq 0. \tag{4.5}$$

We now complete the proof of Theorem 14.

Proof of Theorem 14. Since Q_∞ is initialized with its stationary distribution (see [101]), it follows from the basic properties of the $M/GI/\infty$ queue (see [101]) that $\mathbb{P}(Q_\infty(s) \geq n) = \mathbb{P}(Z \geq n)$, and conditional on the event $\{Q_\infty(s) \geq n\}$, the set of remaining processing times (at time s) of those jobs being processed in $\mathcal{Q}_\infty$ at time s are drawn i.i.d. distributed as $R(S)$. Thus conditional on the event $\{Q_\infty(s) \geq n\}$, one has that $|\mathcal{C}^s| = n$, and the set of remaining processing times (at time s, in $\mathcal{Q}_\infty$) of those jobs belonging to $\mathcal{C}^s$ is drawn i.i.d. distributed as $R(S)$.

By construction the number of jobs initially in $\mathcal{Q}_2^s$ at time 0 *plus* the number of jobs that arrive to $\mathcal{Q}_2^s$ on $(0, s]$ is at most n. Thus all jobs initially in $\mathcal{Q}_2^s$ at time 0 and all jobs that arrive to $\mathcal{Q}_2^s$ on $(0, s]$ begin processing immediately in $\mathcal{Q}_2^s$, as if $\mathcal{Q}_2^s$ were an infinite-server queue. It follows from our construction that conditional on the event $\{Q_\infty(s) \geq n\}$, the set of remaining processing times (at time s) of the n jobs in $\mathcal{Q}_2^s$ at time s equals the set of remaining processing times (at time s, in $\mathcal{Q}_\infty$) of those jobs belonging to $\mathcal{C}^s$, and are thus drawn i.i.d. distributed as $R(S)$.

Let us fix some s, t s.t. $0 \leq s \leq t$. Recall that V_i^j denotes the length of the jth renewal interval in process $N_i(t), j \geq 1, i = 1, \ldots, n$. It follows from our construction that conditional on the event $\{Q_\infty(s) \geq n\}$, we may set the remaining processing time (at time s) of the job on server i in $\mathcal{Q}_2^s$ at time s equal to V_i^1. We can also set the processing time of the jth job assigned to server i in $\mathcal{Q}_2^s$ (after time s) equal to V_i^{j+1}. Under this coupling the total number of jobs that depart from server i in $\mathcal{Q}_2^s$ during $[s, t]$ is at most $N_i(t-s)$, and therefore the total number of departures from $\mathcal{Q}_2^s$ during $[s, t]$ is at most $\sum_{i=1}^n N_i(t-s)$, independent of the arrival process to $\mathcal{Q}_2^s$ on $[s, t]$. By the memoryless and stationary increments properties of the Poisson process, we may let the arrival process to $\mathcal{Q}_2^s$ on $[s, t]$ equal $A(v)_{0 \leq v \leq t-s}$. Combining the above, we find that for all $x > 0$, $\mathbb{P}(Q_2^s(t) - n > x) \geq \mathbb{P}(Z \geq n)\mathbb{P}\big(A(t-s) - \sum_{i=1}^n N_i(t-s) > x\big)$. Observing that s was general, we may then take the supremum of the above bound over all $s \in [0, t]$, and combine with (4.5) to complete the proof of the theorem. The corresponding steady-state result then follows from the fact that monotonic sequences have limits and the definition of weak convergence. □

4.5 Proof of Tightness Result

In this section, we prove Theorem 11. We note that it follows almost immediately from Theorem 13 and well-known tightness results from the literature (see [7] Theorem 14.6, [113] Theorem 7.2.3) that for any *fixed* $T \geq 0$, $\{n^{-\frac{1}{2}}\big(Q^n(t)-n\big)^+_{0\leq t\leq T}, n \geq 1\}$ is tight in the space $D[0,T]$ under the J_1 topology (see Subsection 4.6.1 for details). The challenge is that when analyzing $\{n^{-\frac{1}{2}}\big(Q^n(\infty)-n\big)^+, n \geq 1\}$, one does not have the luxury of bounded time intervals. In particular, to apply Theorem 13, we must show tightness of a supremum taken over an infinite time horizon. For this reason, most standard weak convergence type results and arguments from the literature (see [113]) break down, and cannot immediately be applied. Instead, we will relate the supremum appearing in the r.h.s. of Theorem 13 to the steady-state waiting time in an appropriate $G/D/1$ queue with stationary (as opposed to i.i.d.) inter-arrival times. We will then apply known results from the literature, in particular [100], to show that under the H-W scaling this sequence of steady-state waiting times, properly normalized, is tight.

Suppose that assumptions H-W and T_0 hold. Let $A_n(t) \triangleq A(\lambda_n t)$. In light of Theorem 13, it suffices to prove that $\{n^{-\frac{1}{2}} \sup_{t\geq 0}\big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big), n \geq 1\}$ is tight. Let $\mathsf{A}^0_n(t)$ denote an ordinary renewal process with renewal distribution $A\lambda_n^{-1}$, independent of $\{N_i(t), i = 1,\ldots,n\}$. Note that we may construct $A_n(t)$ and $A^0_n(t)$ on the same probability space so that $A_n(t) \leq 1 + A^0_n(t)$ for all $t \geq 0$. It thus suffices to demonstrate the tightness of $\{n^{-\frac{1}{2}} \sup_{t\geq 0}\big(A^0_n(t) - \sum_{i=1}^{n} N_i(t)\big), n \geq 1\}$.

Let $\{A^1_i, i \geq 1\}$ denote a countably infinite sequence of r.v.s drawn i.i.d. distributed as A, independent of $\{N_i(t), i = 1,\ldots,n\}$. Note that since $A^0_n(t) - \sum_{i=1}^{n} N_i(t)$ only increases at jumps of $A^0_n(t)$, we may construct $A^0_n(t), \sum_{i=1}^{n} N_i(t)$, and $\{A^1_i, i \geq 1\}$

on the same probability space so that

$$n^{-\frac{1}{2}}\sup_{t\geq 0}\big(A^0_n(t)-\sum_{i=1}^{n}N_i(t)\big)=n^{-\frac{1}{2}}\sup_{k\geq 0}\big(k-\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k}A^1_j)\big). \tag{4.6}$$

We now show that

$$\{n^{-\frac{1}{2}}\sup_{k\geq 0}\big(k-\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k}A^1_j)\big), n\geq 1\} \tag{4.7}$$

is tight, which (by the above) will imply Theorem 11. Fortunately, the tightness of such sequences of suprema has already been addressed in the literature, in the context of steady-state waiting times in a $G/G/1$ queue, with stationary inter-arrival times, in heavy-traffic. In particular, note that for $M\geq 1$, $\sup_{0\leq k\leq M}\big(k-\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k}A^1_j)\big)$ corresponds to the waiting time of the $(M+1)$st arrival to a $G/D/1$ queue, initially empty, with all processing times equal to 1, and the kth inter-arrival time equal to

$$\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{M-k}A^1_j,\lambda_n^{-1}\sum_{j=1}^{M-k+1}A^1_j), k\leq M.$$

Recall that $\sum_{i=1}^{n}N_i(t)_{t\geq 0}$ has the same distribution (on the process level) as $\sum_{i=1}^{n}N_i\big(t-s,t\big)_{0\leq s\leq t}$ (see [27]), and $\{A^1_i, i\geq 1\}$ are i.i.d. It follows that for all $M\geq 1$, $\sup_{0\leq k\leq M}\big(k-\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k}A^1_j)\big)$ also has the same distribution as the waiting time of the $(M+1)$st arrival to a $G/D/1$ queue, initially empty, with all processing times equal to 1, and the kth inter-arrival time equal to

$$\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k-1}A^1_j,\lambda_n^{-1}\sum_{j=1}^{k}A^1_j), k\geq 1.$$

For this queueing model, in which the sequence of inter-arrival times is stationary, one can ask whether there is a meaningful notion of steady-state waiting time, whose distribution would naturally coincide with that of

$$\lim_{M\to\infty}\sup_{0\leq k\leq M}\big(k-\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k}A_j^1)\big)=\sup_{k\geq 0}\big(k-\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k}A_j^1)\big).$$

Furthermore, should one examine a sequence of such queues in heavy traffic, one can ask whether the corresponding sequence of steady-state waiting times, properly normalized, is tight.

Note that as (4.7) is such a sequence, we are left to answer exactly this question. Fortunately, sufficient conditions for tightness of such a sequence are given in [100]. In particular, as we will show, it follows from the results of [100] (in the notation of [100]) that

Theorem 15. *Suppose that for all sufficiently large* n, $\{\zeta_{n,i}, i\geq 1\}$ *is a stationary, countably infinite sequence of r.v. Let* $a_n \triangleq \mathbb{E}[\zeta_{n,1}]$, *and* $W_{n,k}\triangleq\sum_{i=1}^{k}\zeta_{n,i}$. *Further assume that* $a_n<0, \lim_{n\to\infty}a_n=0$, *and there exist* $C_1, C_2<\infty$ *and* $\epsilon>0$ *s.t. for all sufficiently large* n,

(i) $\mathbb{E}\big[|W_{n,k}-ka_n|^{2+\epsilon}\big]\leq C_1k^{1+\frac{\epsilon}{2}}$ *for all* $k\geq 1$;

(ii) $\mathbb{P}\big(\max_{i=1,\dots,k}(W_{n,i}-ia_n)>x\big)\leq C_2k^{1+\frac{\epsilon}{2}}x^{-(2+\epsilon)}$ *for all* $k\geq 1$ *and* $x>0$.

Then $\{|a_n|\sup_{k\geq 0}W_{n,k}, n\geq 1\}$ *is tight.*

Proof. The proof follows from Theorem 1 of [100], and is deferred to the appendix. □

To verify that the assumptions of Theorem 15 hold for

$$\{n^{-\frac{1}{2}}\sup_{k\geq 0}\big(k-\sum_{i=1}^{n}N_i(\lambda_n^{-1}\sum_{j=1}^{k}A_j^1)\big), n\geq 1\},$$

we will rely on a technical result from [7], which gives a bound on the supremum of a general random walk in terms of bounds on its increments. In particular, it is shown in [7] Theorem 10.2 that

Lemma 27. *Suppose $k < \infty$, $X_1, X_2, \ldots, X_k$ is a sequence of general (possibly dependent and not identically distributed) random variables, $S_j \triangleq \sum_{i=1}^{j} X_i$, and $M_k = \max_{j \leq k} |S_j|$. Further suppose that there exist real numbers $\alpha > \frac{1}{2}$, $\beta \geq 0$, and a sequence of non-negative numbers $u_1, u_2, \ldots, u_k$ s.t. for all $0 \leq i \leq j \leq k$ and $x > 0$,*

$$\mathbb{P}\big(|S_j - S_i| \geq x\big) \leq x^{-4\beta}\big(\sum_{i<l\leq j} u_l\big)^{2\alpha}.$$

Then there exists a finite constant $K_{\alpha,\beta}$, depending only on α and β, s.t. for all $x > 0$,

$$\mathbb{P}\big(M_k \geq x\big) \leq K_{\alpha,\beta} x^{-4\beta}\big(\sum_{0<l\leq k} u_l\big)^{2\alpha}.$$

We will also use frequently the inequality

$$(x_1 + x_2)^r \leq 2^{r-1}x_1^r + 2^{r-1}x_2^r \text{ for all } r \geq 1 \text{ and } x_1, x_2 \geq 0, \tag{4.8}$$

which follows from the convexity of $f(x) \triangleq x^r$, $r \geq 1$.

Before proceeding with the proof of Theorem 11, we establish two more auxiliary results. The first bounds the moments of the sum of n i.i.d. zero-mean r.v. in terms of the moments of the individual r.v.s. and n, and is proven in [114].

Lemma 28. *For all $r \geq 2$, there exists $C_r < \infty$ (depending only on r) s.t. for all r.v. X satisfying $\mathbb{E}[X] = 0$ and $\mathbb{E}[|X|^r] < \infty$, if $\{X_i, i \geq 1\}$ is an i.i.d. sequence of r.v.s distributed as X, then for all $k \geq 1$,*

$$\mathbb{E}[|\sum_{i=1}^{k} X_i|^r] \leq C_r k^{\frac{r}{2}} \mathbb{E}[|X|^r].$$

Second, we prove a bound for the central moments of a pooled equilibrium renewal process.

Lemma 29. *Let X denote any non-negative r.v. s.t.* $\mathbb{E}[X] = \mu^{-1} \in (0, \infty)$, *and* $\mathbb{E}[X^r] < \infty$ *for some* $r \geq 2$. *Let* $\{Z_i^e(t), i \geq 1\}$ *denote a set of i.i.d. equilibrium renewal processes with renewal distribution X. Then there exists* $C_{X,r} < \infty$ *(depending only on X and r) s.t. for all* $n \geq 1$ *and* $t \geq 0$,

$$\mathbb{E}\big[|\sum_{i=1}^{n} Z_i^e(t) - \mu n t|^r\big] \leq C_{X,r}\big(1 + (nt)^{\frac{r}{2}}\big). \tag{4.9}$$

Proof. The proof is deferred to the appendix. □

With the above bounds at our disposal, we now complete the proof of Theorem 11.

Proof of Theorem 11. In the notation of Theorem 15, let

$$\zeta_{n,k} \overset{\Delta}{=} 1 - \sum_{i=1}^{n} N_i(\lambda_n^{-1} \sum_{j=1}^{k-1} A_j^1, \lambda_n^{-1} \sum_{j=1}^{k} A_j^1),$$

$$W_{n,k} \overset{\Delta}{=} k - \sum_{i=1}^{n} N_i(\lambda_n^{-1} \sum_{j=1}^{k} A_j^1).$$

That $\{\zeta_{n,i}, i \geq 1\}$ is a stationary, countably infinite sequence of r.v. follows from the stationary increments property of the equilibrium renewal process. Since $\mathbb{E}[\sum_{i=1}^{n} N_i(t)] = nt\mu$ for all $t \geq 0$, it follows that $a_n \overset{\Delta}{=} \mathbb{E}[\zeta_{n,1}] = 1 - \frac{n}{\lambda_n} = -\frac{B}{n^{\frac{1}{2}} - B} < 0$, and $\lim_{n \to \infty} a_n = 0$. Thus we need only verify assumptions (i) and (ii) of Theorem 15. Since $\mathbb{E}[A^{2+\epsilon}], \mathbb{E}[S^{2+\epsilon}] < \infty$ for some $\epsilon > 0$ by the T_0 assumptions, we may fix some $r > 2$ s.t. $\mathbb{E}[A^r], \mathbb{E}[S^r] < \infty$. Note that $\mathbb{E}\big[|W_{n,k} - ka_n|^r\big]$ equals

$$\begin{aligned}
& \mathbb{E}[|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \frac{kn}{\lambda_n}|^r] \\
\leq \quad & \mathbb{E}\Big[\Big(|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1| + |\mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1 - \frac{kn}{\lambda_n}|\Big)^r\Big] \\
\leq \quad & 2^{r-1}\mathbb{E}[|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1|^r] \qquad (4.10) \\
& +2^{r-1}\mathbb{E}[|\mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1 - \frac{kn}{\lambda_n}|^r] \quad \text{by (4.8).} \qquad (4.11)
\end{aligned}$$

We now bound (4.10). By Lemmas 28 - 29, there exist $C_{S,r}, C_r < \infty$ independent of n and k s.t. $\mathbb{E}\big[|\sum_{i=1}^{n} N_i(\lambda_n^{-1}\sum_{j=1}^{k} A_j^1) - \mu\frac{n}{\lambda_n}\sum_{j=1}^{k} A_j^1|^r\big]$ is at most

$$\begin{aligned}
& C_{S,r} + C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\mathbb{E}[(\sum_{j=1}^{k} A_j^1)^{\frac{r}{2}}] \quad \text{by Lemma 29} \\
& \leq C_{S,r} + C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\mathbb{E}\Big[\Big(|\sum_{j=1}^{k}(A_j^1 - \mu^{-1})| + k\mu^{-1}\Big)^{\frac{r}{2}}\Big] \\
& \leq C_{S,r} + C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\Big(2^{\frac{r}{2}-1}\mathbb{E}[|\sum_{j=1}^{k}(A_j^1 - \mu^{-1})|^{\frac{r}{2}}] + 2^{\frac{r}{2}-1}(k\mu^{-1})^{\frac{r}{2}}\Big) \\
& \leq C_{S,r} + 2^{\frac{r}{2}-1}C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\Big(\mathbb{E}^{\frac{1}{2}}[|\sum_{j=1}^{k}(A_j^1 - \mu^{-1})|^r] + (k\mu^{-1})^{\frac{r}{2}}\Big) \\
& \qquad \text{since } \mathbb{E}[X] \leq \mathbb{E}^{\frac{1}{2}}[X^2] \text{ for any non-negative r.v. } X \\
& \leq C_{S,r} + 2^{\frac{r}{2}-1}C_{S,r}(\frac{n}{\lambda_n})^{\frac{r}{2}}\Big(\big(C_r k^{\frac{r}{2}}\mathbb{E}[|A - \mu^{-1}|^r]\big)^{\frac{1}{2}} + (k\mu^{-1})^{\frac{r}{2}}\Big) \\
& \qquad \text{by Lemma 28.} \\
& \leq C_1' k^{\frac{r}{2}}, \qquad (4.12)
\end{aligned}$$

for some finite constant C_1' independent of n and k, since $\mathbb{E}\big[|A - \mu^{-1}|^r\big] < \infty$, and

$\lim_{n\to\infty}\frac{n}{\lambda_n}=1$.

We now bound (4.11).

$$\begin{aligned}
\mathbb{E}\big[|\mu\frac{n}{\lambda_n}\sum_{j=1}^{k}A_j^1-\frac{kn}{\lambda_n}|^r\big] &= (\frac{n}{\lambda_n})^r\mu^r\mathbb{E}\big[|\sum_{j=1}^{k}(A_j^1-\mu^{-1})|^r\big] \\
&\leq \Big(C_r(\frac{n}{\lambda_n})^r\mu^r\mathbb{E}\big[|A-\mu^{-1}|^r\big]\Big)k^{\frac{r}{2}} \text{ by Lemma 28} \\
&\leq C_1''k^{\frac{r}{2}}, \qquad (4.13)
\end{aligned}$$

for some finite constant C_1'', independent of n and k. Using (4.12) to bound (4.10) and (4.13) to bound (4.11), it follows that assumption (i) of Theorem 15 holds for the finite constant $C_1 \triangleq 2^{r-1}(C_1'+C_1'')$. We now apply Lemma 27 to show that assumption (ii) holds as well. In the notation of Lemma 27, let $S_{n,i}\triangleq W_{n,i}-ia_n$ for $i\geq 0$, and $M_{n,k}\triangleq \max_{i\leq k}|W_{n,i}-ia_n|$ for $k\geq 0$. Then for all n, $0\leq i\leq j$, and $x>0$,

$$\begin{aligned}
\mathbb{P}\big(|S_{n,j}-S_{n,i}|\geq x\big) &= \mathbb{P}\big(|S_{n,j-i}|\geq x\big) \text{ by stationary increments} \\
&= \mathbb{P}(|W_{n,j-i}-(j-i)a_n|\geq x) \\
&\leq C_1(j-i)^{\frac{r}{2}}x^{-r} \text{ by Markov's inequality} \\
&\leq \big((C_1+1)(j-i)\big)^{\frac{r}{2}}x^{-r}.
\end{aligned}$$

Thus for all n and $k\geq 1$, we may apply Lemma 27 (in the notation of Lemma 27) with $\beta\triangleq\frac{r}{4},\alpha\triangleq\frac{r}{4}$, and $u_l\triangleq(C_1+1)$ for $1\leq l\leq k$, to find that there exists a constant $K_r<\infty$ (depending only on r) s.t. for all $x>0$,

$$\mathbb{P}\big(\max_{i=1,\ldots,k}(W_{n,i}-ia_n)>x\big)\leq K_r(C_1+1)^{\frac{r}{2}}k^{\frac{r}{2}}x^{-r}. \qquad (4.14)$$

It follows that assumption (ii) of Theorem 15 holds as well, with (in the notation of Theorem 15) $C_2\triangleq K_r(C_1+1)^{\frac{r}{2}}$, $\epsilon\triangleq r-2$. Combining the above, we find that all assumptions of Theorem 15 hold, and thus we may apply Theorem 15 to find that

$$\left\{\frac{B}{n^{\frac{1}{2}} - B} \sup_{k \geq 0} \big(k - \sum_{i=1}^{n} N_i(\lambda_n^{-1} \sum_{j=1}^{k} A_j^1)\big), n \geq 1\right\}$$

is tight. Combining with (4.6) completes the proof of Theorem 11. □

4.6 Proof of Large Deviations Results

In this section, we complete the proofs of our main results. We proceed by combining our upper and lower bounds with several known weak convergence results for (pooled) renewal processes and the suprema of Gaussian processes. Recall that a Gaussian process on $\mathbb{R}$ is a stochastic process $Z(t)_{t \geq 0}$ s.t. for any finite set of times $t_1, \ldots, t_k$, the vector $\big(Z(t_1), \ldots, Z(t_k)\big)$ has a Gaussian distribution. A Gaussian process $Z(t)$ is known to be uniquely determined by its mean function $\mathbb{E}[Z(t)]$ and covariance function $\mathbb{E}[Z(s)Z(t)]$, and refer the reader to [35],[55],[2],[77], and the references therein for details on existence, continuity, etc.

4.6.1 Preliminary weak convergence results

In this subsection we review several weak convergence results for renewal processes, and apply them to $A_n(t)$ and $\sum_{i=1}^{n} N_i(t)$. For an excellent review of weak convergence, and the associated spaces (e.g. $D[0, T]$) and topologies/metrics (e.g. uniform, J_1), the reader is referred to [113]. Let $\mathcal{A}(t)$ denote the w.p.1 continuous Gaussian process s.t. $\mathbb{E}[\mathcal{A}(t)] = 0, \mathbb{E}[\mathcal{A}(s)\mathcal{A}(t)] = \mu c_A^2 \min(s, t)$, namely $\mathcal{A}(t)$ is a driftless Brownian motion. Then it follows from the well-known Functional Central Limit Theorem (FCLT) for renewal processes (see [7] Theorem 14.6) that

Theorem 16. *For any $T \in [0, \infty)$, the sequence of processes*

$\{\lambda_n^{-\frac{1}{2}}\big(A_n(t) - \lambda_n\mu t\big)_{0\le t\le T}, n \ge 1\}$ converges weakly to $\mathcal{A}(t)_{0\le t\le T}$ in the space $D[0,T]$ under the J_1 topology.

We now give a weak convergence result for $\sum_{i=1}^n N_i(t)$, which is stated in [113] (see Theorem 7.2.3) and formally proven in [110] (see Theorem 2).

Theorem 17. *There exists a w.p.1 continuous Gaussian process $\mathcal{D}(t)$ s.t. $\mathbb{E}[\mathcal{D}(t)] = 0, \mathbb{E}[\mathcal{D}(s)\mathcal{D}(t)] = \mathbb{E}[\big(N_1(s) - \mu s\big)\big(N_1(t) - \mu t\big)]$ for all $s, t \ge 0$. Furthermore, for any $T \in [0,\infty)$, the sequence of processes $\{n^{-\frac{1}{2}}\big(\sum_{i=1}^n N_i(t) - n\mu t\big)_{0\le t\le T}, n \ge 1\}$ converges weakly to $\mathcal{D}(t)_{0\le t\le T}$ in the space $D[0,T]$ under the J_1 topology.*

We note that the T_0 assumptions (i) and (iii), which guarantee that $\mathbb{E}[S^{2+\epsilon}] < \infty$ and $\limsup_{x\downarrow 0} x^{-1}\mathbb{P}(S \le x) < \infty$, ensure that the technical conditions required to apply [113] Theorem 7.2.3, namely that $E[S^2] < \infty$ and $\limsup_{x\downarrow 0} x^{-1}\big(\mathbb{P}(S \le x) - \mathbb{P}(S = 0)\big) < \infty$, hold.

It follows from Theorems 16 - 17 that

Lemma 30. *For any fixed $T \ge 0$, $\{n^{-\frac{1}{2}}\big(A_n(t) - \sum_{i=1}^n N_i(t)\big)_{0\le t\le T}, n \ge 1\}$ converges weakly to $\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big)_{0\le t\le T}$ in the space $D[0,T]$ under the J_1 topology.*

Proof. Note that $n^{-\frac{1}{2}}\big(A_n(t) - \sum_{i=1}^n N_i(t)\big)_{0\le t\le T}$ equals

$$\left(\lambda_n^{\frac{1}{2}} n^{-\frac{1}{2}}\big(A_n(t) - \lambda_n\mu t\big)\lambda_n^{-\frac{1}{2}} - \big(\sum_{i=1}^n N_i(t) - n\mu t\big)n^{-\frac{1}{2}} - B\mu t\right)_{0\le t\le T}.$$

The lemma then follows from Theorems 16 - 17. □

We note that a process very similar to $\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big)_{0\le t\le T}$ was studied in [110] as the weak limit of a sequence of queues with superposition arrival processes. The continuity of the supremum map in the space $D[0,T]$ under the J_1 topology (see [113] Theorem 13.4.1), combined with Lemma 30, implies that

Corollary 4. *For any fixed* $T \geq 0$, $\{n^{-\frac{1}{2}} \sup_{0 \leq t \leq T} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big), n \geq 1\}$ *converges weakly to the r.v.* $\sup_{0 \leq t \leq T} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big)$.

4.6.2 Preliminary large deviation results

Before proceeding with the remaining proofs, we will need to establish some results from the theory of large deviations of Gaussian processes and their suprema. We note that the relationship between the suprema of Gaussian processes and queueing systems is well known (see [36]). We will rely heavily on the following theorem, proven in Section 3.1 of [36].

Theorem 18. *Suppose* $\mathcal{Z}(t)$ *is a Gaussian process with stationary increments s.t.* $\mathbb{E}[\mathcal{Z}(t)] = 0$ *for all* $t \geq 0$, *and* $\lim_{t \to \infty} t^{-1}\mathbb{E}[\mathcal{Z}^2(t)] = \sigma^2 > 0$. *Then for any* $c > 0$,

$$\lim_{x \to \infty} x^{-1} \log \bigg(\mathbb{P}\big(\sup_{t \geq 0}(\mathcal{Z}(t) - ct) > x\big) \bigg) = -\frac{2c}{\sigma^2}.$$

It is also implicit from [36] (although we include a short proof) that

Theorem 19. *Under the same assumptions as Theorem 18, for any* $c > 0$,

$$\lim_{x \to \infty} x^{-1} \log \bigg(\sup_{t \geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big) \bigg) = -\frac{2c}{\sigma^2}.$$

Proof. That $\limsup_{x \to \infty} x^{-1} \log \bigg(\sup_{t \geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big) \bigg) \leq -\frac{2c}{\sigma^2}$ follows immediately from Theorem 18 and the fact that $\sup_{t \geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big) \leq \mathbb{P}\bigg(\sup_{t \geq 0} \big(\mathcal{Z}(t) - ct\big) > x \bigg)$.

Letting $t = \frac{x}{c}$, we find that

$$\sup_{t \geq 0} \mathbb{P}\big(\mathcal{Z}(t) - ct > x\big) \geq \mathbb{P}\big(\mathcal{Z}(\frac{x}{c}) - x > x\big). \tag{4.15}$$

Let G denote a normally distributed r.v. with mean 0 and variance 1. Then since $\mathcal{Z}(\frac{x}{c})$ is normally distributed with mean zero, it follows from (4.15) that

$$\sup_{t\geq 0}\mathbb{P}\big(\mathcal{Z}(t)-ct>x\big)\geq \mathbb{P}\bigg(G>2x\mathbb{E}^{-\frac{1}{2}}\big[\mathcal{Z}^2(\frac{x}{c})\big]\bigg). \tag{4.16}$$

We use the following identity from [1] Equation 7.1.13. Namely, for all $y>0$,

$$\mathbb{P}(G>y)\geq \big(y+(y^2+4)^{-\frac{1}{2}}\big)^{-1}(\frac{2}{\pi})^{\frac{1}{2}}\exp(-\frac{y^2}{2}).$$

Thus

$$\mathbb{P}(G>y)\geq \exp(-\frac{y^2}{2}-y) \text{ for all sufficiently large } y. \tag{4.17}$$

By assumption, $\lim_{t\to\infty}t^{-1}\mathbb{E}[\mathcal{Z}^2(t)]=\sigma^2>0$, and thus $\lim_{x\to\infty}2x\mathbb{E}^{-\frac{1}{2}}\big[\mathcal{Z}^2(\frac{x}{c})\big]=\infty$. It thus follows from (4.16) and (4.17) that for all sufficiently large x,

$$x^{-1}\log\bigg(\sup_{t\geq 0}\mathbb{P}\big(\mathcal{Z}(t)-ct>x\big)\bigg)\geq -2x\mathbb{E}^{-1}\big[\mathcal{Z}^2(\frac{x}{c})\big]-2\mathbb{E}^{-\frac{1}{2}}\big[\mathcal{Z}^2(\frac{x}{c})\big].$$

Since $\lim_{x\to\infty}(\frac{x}{c})^{-1}\mathbb{E}[\mathcal{Z}^2(\frac{x}{c})]=\sigma^2$, it follows that $\liminf_{x\to\infty}x^{-1}\log\bigg(\sup_{t\geq 0}\mathbb{P}\big(\mathcal{Z}(t)-ct>x\big)\bigg)\geq -\frac{2c}{\sigma^2}$, concluding the proof of the theorem. □

In light of Theorem 18, Theorem 19 can be interpreted as saying that such a process is 'most likely' to exceed a given value x at a particular time (roughly $\frac{x}{c}$), and much less likely to exceed that value at any other time (see the discussion in [36]). We note that the duality of Theorems 18 - 19 coincides with the duality exhibited by our upper and lower bounds (Theorems 13 - 14) - a relationship that we will exploit to prove our large deviation results.

We are now in a position to apply Theorems 18 - 19 to $\mathcal{A}(t)-\mathcal{D}(t)$.

Corollary 5.

(i) $\lim_{x\to\infty} x^{-1} \log \mathbb{P}\Big(\sup_{t\ge 0} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) > x\Big) = -2B(c_A^2 + c_S^2)^{-1}$;

(ii) $\lim_{x\to\infty} x^{-1} \log \Big(\sup_{t\ge 0} \mathbb{P}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t > x\big)\Big) = -2B(c_A^2 + c_S^2)^{-1}$.

Proof. That $\mathcal{A}(t) - \mathcal{D}(t)$ is a zero-mean Gaussian process with stationary increments follows from definitions, the independence of $\mathcal{A}(t)$ and $\mathcal{D}(t)$, and the fact that both $\mathcal{A}(t)$ and $N_1(t)$ have stationary increments. Note that

$$\mathbb{E}[\big(\mathcal{A}(t) - \mathcal{D}(t)\big)^2] = \mu c_A^2 t + \mathbb{E}[\big(N_1(t) - \mu t\big)^2]. \tag{4.18}$$

We claim that $\lim_{t\to\infty} t^{-1}\mathbb{E}[\big(N_1(t) - \mu t\big)^2] = \mu c_S^2$. Indeed, let G_S denote a normally distributed r.v. with mean 0 and variance μc_S^2. It follows from the well-known Central Limit Theorem for renewal processes (see [93] Theorem 3.3.5), and the fact that $h(z) \triangleq z^2$ is a continuous function, that the sequence of r.v.s $\{\Big(t^{-\frac{1}{2}}\big(N_1(t) - \mu t\big)\Big)^2, t \ge 1\}$ converges weakly to G_S^2. Recall that $\mathbb{E}[S^{2+\epsilon}] < \infty$ for some $\epsilon > 0$ by the T_0 assumptions. Thus it follows from Lemma 29 that the sequence of r.v.s $\{\Big(t^{-\frac{1}{2}}\big(N_1(t) - \mu t\big)\Big)^2, t \ge 1\}$ is uniformly integrable. It follows that $\lim_{t\to\infty} t^{-1}\mathbb{E}[\big(N_1(t) - \mu t\big)^2] = \mu c_S^2$, since uniform integrability plus weak convergence implies convergence of moments.

Combining with (4.18), we find that $\lim_{t\to\infty} t^{-1}\mathbb{E}[\big(\mathcal{A}(t) - \mathcal{D}(t)\big)^2] = \mu(c_A^2 + c_S^2) > 0$ by the T_0 assumptions. It follows that $\mathcal{A}(t) - \mathcal{D}(t)$ satisfies the conditions needed to apply Theorems 18 - 19, from which the corollary follows. □

4.6.3 Proof of Theorem 12

Before completing the proofs of our main results, it will be useful to prove a strengthening of Theorem 11. Namely,

Lemma 31. *For all $x \geq 0$,*

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mathbb{P}\Bigg(n^{-\frac{1}{2}} \sup_{t\geq T} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) > x \Bigg) = 0. \tag{4.19}$$

Proof. Since $\mathbb{E}[A^{2+\epsilon}], \mathbb{E}[S^{2+\epsilon}] < \infty$ for some $\epsilon > 0$ by the T_0 assumptions, we may fix some $r > 2$ s.t. $\mathbb{E}[A^r], \mathbb{E}[S^r] < \infty$. Note that since $x \geq 0$, $\mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t\geq T} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) > x \Big)$ is at most $\mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t\geq T} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) > 0 \Big)$. By a simple union bound, $\mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t\geq T} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) > 0 \Big)$ is at most

$$\mathbb{P}\Bigg(n^{-\frac{1}{2}} \big(A_n(T) - \sum_{i=1}^{n} N_i(T)\big) > -\frac{B}{2}\mu T \Bigg) \tag{4.20}$$

$$+\mathbb{P}\Bigg(\sup_{t\geq T} \Bigg(n^{-\frac{1}{2}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) - n^{-\frac{1}{2}} \big(A_n(T) - \sum_{i=1}^{n} N_i(T)\big) \Bigg) > \frac{B}{2}\mu T \Bigg) \tag{4.21}$$

We now bound (4.20), which equals

$$\begin{aligned}
& \mathbb{P}\Bigg(n^{-\frac{1}{2}} \big(A_n(T) - \lambda_n \mu T\big) - n^{-\frac{1}{2}} \big(\sum_{i=1}^{n} N_i(T) - n\mu T\big) - B\mu T > -\frac{B}{2}\mu T \Bigg) \\
\leq \quad & \mathbb{P}\Bigg(|A_n(T) - \lambda_n \mu T| + |\sum_{i=1}^{n} N_i(T) - n\mu T| > n^{\frac{1}{2}} \frac{B}{2}\mu T \Bigg) \text{ by the tri. ineq.} \\
\leq \quad & 2^{r-1} \Bigg(\mathbb{E}\big[|A_n(T) - \lambda_n \mu T|^r\big] + \mathbb{E}\big[|\sum_{i=1}^{n} N_i(T) - n\mu T|^r\big] \Bigg) n^{-\frac{r}{2}} \big(\frac{B}{2}\mu T\big)^{-r} \qquad (4.22) \\
& \text{by Markov's inequality and (4.8).}
\end{aligned}$$

W.l.o.g. assuming $nT \geq \lambda_n T \geq 1$, it follows from Lemma 29 (applied with $n = 1$), and the fact that $A_n(T)$ has the same distribution as $A(\lambda_n T)$, that there exists $C_{A,r} \triangleq \sup_{t\geq 1} t^{-\frac{r}{2}}\mathbb{E}\big[|A(t)-\mu t|^r\big] < \infty$ s.t.

$$\mathbb{E}\big[|A_n(T)-\lambda_n\mu T|^r\big] \leq C_{A,r}(\lambda_n T)^{\frac{r}{2}} \leq C_{A,r}(nT)^{\frac{r}{2}}. \tag{4.23}$$

Since $nT \geq 1$ by assumption, it follows from Lemma 29 that there exist $C_{S,r} < \infty$ s.t.

$$\mathbb{E}\big[|\sum_{i=1}^{n} N_i(T) - n\mu T|^r\big] \leq C_{S,r}(nT)^{\frac{r}{2}}. \tag{4.24}$$

It follows from (4.23) and (4.24) that (4.22) is at most

$$2^{r-1}(C_{A,r}+C_{S,r})(\frac{B}{2}\mu)^{-r}T^{-\frac{r}{2}}.$$

Thus we find that

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}\big(A_n(T)-\sum_{i=1}^{n}N_i(T)\big) > -\frac{B}{2}\mu T\Big) = 0. \tag{4.25}$$

We now bound (4.21), which equals $\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq 0}\big(A_n(t)-\sum_{i=1}^{n}N_i(t)\big) > \frac{B}{2}\mu T\Big)$ by stationary increments. But as our proof of Theorem 11 demonstrates tightness of $\{n^{-\frac{1}{2}}\sup_{t\geq 0}\big(A_n(t)-\sum_{i=1}^{n}N_i(t)\big), n\geq 1\}$, it follows that

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq 0}\big(A_n(t)-\sum_{i=1}^{n}N_i(t)\big) > \frac{B}{2}\mu T\Big) = 0. \tag{4.26}$$

Using (4.26) to bound (4.21), we find that

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{t\geq T}\Big(n^{-\frac{1}{2}}\big(A_n(t)-\sum_{i=1}^{n}N_i(t)\big)-n^{-\frac{1}{2}}\big(A_n(T)-\sum_{i=1}^{n}N_i(T)\big)\Big) > \frac{B}{2}\mu T\Big) = 0. \tag{4.27}$$

Combining (4.25) and (4.27) completes the proof. □

We now complete the proof of Theorem 12.

Proof. We first prove the upper bound. By Lemma 31, for any $x > 0$, we may construct a strictly increasing sequence of integers $\{T_{x,k^{-1}}, k \geq 1\}$ s.t. for all $k \geq 1$,

$$\limsup_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t \geq T_{x.k^{-1}}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) \geq x \Big) < k^{-1}.$$

It follows that for all $x > 0$ and $k \geq 1$,

$$\limsup_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t \geq 0} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) \geq x \Big) \tag{4.28}$$

$$\leq \limsup_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{0 \leq t \leq T_{x.k^{-1}}} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) \geq x \Big) + k^{-1}.$$

By the Portmanteau Theorem (see [7]), a sequence of r.v.s $\{X_n\}$ converges weakly to the r.v. X_∞ iff for all closed subsets C of $\mathbb{R}$, $\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \leq \mathbb{P}(X_\infty \in C)$ iff for all open subsets O of $\mathbb{R}$, $\mathbb{P}(X_\infty \in O) \leq \liminf_{n\to\infty} \mathbb{P}(X_n \in O)$. It follows from (4.28) and Corollary 4 that for all $x > 0$ and $k \geq 1$,

$$\limsup_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{t \geq 0} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) \geq x \Big) \tag{4.29}$$

$$\leq \mathbb{P}\Big(\sup_{0 \leq t \leq T_{x.k^{-1}}} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) \geq x \Big) + k^{-1}.$$

Note that the sequence of events $\Big\{ \sup_{0 \leq t \leq T_{x.k^{-1}}} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) \geq x, k \geq 1 \Big\}$ is monotone in k. It follows that

$$\lim_{k\to\infty} \mathbb{P}\Bigg(\sup_{0\le t\le T_{x.k^{-1}}} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) \ge x \Bigg) = \mathbb{P}\Bigg(\sup_{t\ge 0} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) \ge x \Bigg).$$

It then follows from (4.29), by letting $k \to \infty$, that for all $x > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\Bigg(n^{-\frac{1}{2}} \sup_{t\ge 0} \big(A_n(t) - \sum_{i=1}^{n} N_i(t)\big) \ge x \Bigg) \le \mathbb{P}\Bigg(\sup_{t\ge 0} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) \ge x \Bigg). \tag{4.30}$$

From Theorem 13 and (4.30) we have

$$\begin{aligned}
& \limsup_{x\to\infty} x^{-1} \log \Bigg(\limsup_{n\to\infty} \mathbb{P}\Bigg(\big(Q^n(\infty) - n\big)^+ n^{-\frac{1}{2}} > x \Bigg) \Bigg) \\
\le\ & \limsup_{x\to\infty} x^{-1} \log \mathbb{P}\Bigg(\sup_{t\ge 0} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) > x - 1 \Bigg) \\
=\ & \limsup_{x\to\infty} \frac{x-1}{x} \Bigg((x-1)^{-1} \log \mathbb{P}\Bigg(\sup_{t\ge 0} \big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) > x - 1 \Bigg) \Bigg) \\
=\ & -2B(c_A^2 + c_S^2)^{-1} \text{ by Corollary 5.(i),}
\end{aligned}$$

which completes the proof of the upper bound.

We now complete the proof of Theorem 12 by demonstrating that if A is an exponentially distributed r.v., then

$$\liminf_{x\to\infty} x^{-1} \log \Bigg(\liminf_{n\to\infty} \mathbb{P}\Bigg(\big(Q^n(\infty) - n\big)^+ n^{-\frac{1}{2}} > x \Bigg) \Bigg) \ge -2B(c_A^2 + c_S^2)^{-1}. \tag{4.31}$$

Let Z_n denote a Poisson r.v. with mean λ_n. It follows from Theorem 14 that for all

$x > 0$, $\liminf_{n\to\infty} \mathbb{P}\Big(\big(\mathrm{Q}^n(\infty) - n\big)^+ n^{-\frac{1}{2}} > x \Big)$ is at least

$$\Big(\liminf_{n\to\infty} \mathbb{P}(Z_n \geq n) \Big) \Big(\liminf_{n\to\infty} \sup_{t\geq 0} \mathbb{P}\big(A_n(t) - \sum_{i=1}^{n} N_i(t) > x\big) \Big). \tag{4.32}$$

Recall that G is a normally distributed r.v. with mean 0 and variance 1. Thus by the Central Limit Theorem,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \geq n) = \mathbb{P}\big(G \geq B\big). \tag{4.33}$$

Note that for any fixed t, $\mathcal{A}(t) - \mathcal{D}(t) - B\mu t$ is a non-degenerate Gaussian r.v., and every $x \in \mathbb{R}$ is a continuity point of the distribution of any non-degenerate Gaussian r.v. It follows from Lemma 30 and the definition of weak convergence that for any fixed $t \geq 0$ and all $x > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(A_n(t) - \sum_{i=1}^{n} N_i(t) > x\big) = \mathbb{P}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t > x\big).$$

Thus for any fixed $x > 0$ and $s \geq 0$,

$$\begin{aligned} \liminf_{n\to\infty} \sup_{t\geq 0} \mathbb{P}\big(A_n(t) - \sum_{i=1}^{n} N_i(t) > x\big) &\geq \liminf_{n\to\infty} \mathbb{P}\big(A_n(s) - \sum_{i=1}^{n} N_i(s) > x\big) \\ &= \mathbb{P}\big(\mathcal{A}(s) - \mathcal{D}(s) - B\mu s > x\big). \end{aligned} \tag{4.34}$$

By fixing $x > 0$ and taking the supremum over all $s \geq 0$ in (4.34), we find that for all $x > 0$,

$$\liminf_{n\to\infty} \sup_{t\geq 0} \mathbb{P}\big(A_n(t) - \sum_{i=1}^{n} N_i(t) > x\big) \geq \sup_{t\geq 0} \mathbb{P}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t > x\big). \tag{4.35}$$

Combining (4.32), (4.33), and (4.35), we find that the l.h.s. of (4.32) is at least

$$\mathbb{P}\big(G \geq B\big) \sup_{t \geq 0} \mathbb{P}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t > x\big). \tag{4.36}$$

(4.31) then follows from (4.36) and Corollary 5.ii. Combining (4.31) with the first part of Theorem 12, which we have already proven, completes the proof. □

4.7 Application to Reed's Weak Limit

In [91], J. Reed resolved the long-standing open question, originally posed in [52], of the tightness and weak convergence for the queue length of the transient $GI/GI/n$ queue in the H-W regime, for a restricted class of initial conditions. However, the associated weak limit is only described implicitly, as the solution to a certain stochastic convolution equation (see [91]). Prior to this thesis, very little was understood about this limiting process.

In this section we derive the first non-trivial bounds for the weak limit of the transient $GI/GI/n$ queue in the H-W regime. Let $\mathcal{Q}^n_1$ denote the FCFS $GI/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as $A\lambda_n^{-1}$, processing times drawn i.i.d. distributed as S, and the following initial conditions. For $i = 1, \ldots, n$, there is a single job initially being processed on server i, and the set of initial processing times of these n initial jobs is drawn i.i.d. distributed as $R(S)$; there are zero jobs waiting in queue, and the first inter-arrival time is distributed as $R(A\lambda_n^{-1})$, independent of the initial processing times of those jobs initially in system. Let $\hat{Q}_1(t)$ denote the unique strong solution to the stochastic convolution equation given in [91] Equation 1.1. Then letting $Q^n_1(t)$ denote the number in system at time t in $\mathcal{Q}^n_1$, it is proven in [91] that

Theorem 20. *For all $T \in (0, \infty)$, the sequence of stochastic processes $\{n^{-\frac{1}{2}}(Q^n_1(t) -$*

$n)^+_{0\le t\le T}, n \ge 1\}$ converges weakly to $\hat{Q}_1(t)_{0\le t\le T}$ in the space $D[0,T]$ under the J_1 topology.

We now apply Theorem 13 to derive the first non-trivial bounds for $\hat{Q}_1(t)$, proving that

Theorem 21. *For all $x > 0$ and $t \ge 0$,*

$$\mathbb{P}(\hat{Q}_1(t) > x) \le \mathbb{P}\Big(\sup_{0\le s\le t} \big(\mathcal{A}(s) - \mathcal{D}(s) - B\mu s\big) \ge x\Big).$$

Proof. Note that we may let the arrival process to Q_1^n be $A_n(t)$. Thus by Theorem 13, for all $x > 0$ and $t \ge 0$,

$$\begin{aligned}
\liminf_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}}(\mathrm{Q}_1^n(t) - n)^+ > x\Big) &\le \liminf_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{0\le s\le t} \big(A_n(s) - \sum_{i=1}^{n} N_i(s)\big) > x\Big) \\
&\le \limsup_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}} \sup_{0\le s\le t} \big(A_n(s) - \sum_{i=1}^{n} N_i(s)\big) \ge x\Big) \\
&\le \mathbb{P}\Big(\sup_{0\le s\le t} \big(\mathcal{A}(s) - \mathcal{D}(s) - B\mu s\big) \ge x\Big), \qquad (4.37)
\end{aligned}$$

by the Portmanteau Theorem. Again applying the Portmanteau Theorem, it follows from Theorem 20 that for all $x > 0$,

$$\mathbb{P}(\hat{Q}_1(t) > x) \le \liminf_{n\to\infty} \mathbb{P}\Big(n^{-\frac{1}{2}}(Q_1^n(t) - n)^+ > x\Big). \qquad (4.38)$$

Combining (4.37) and (4.38) completes the proof. □

Theorem 21 implies that $\hat{Q}_1(t)$ is distributionally bounded over time, and thus in a sense stable. In particular, for all $t \ge 0$, $\hat{Q}_1(t)$ is stochastically dominated by the r.v. $\sup_{t\ge 0}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big)$. Prior to this thesis, the tightness of $\hat{Q}_1(t)$ as $t \to \infty$ was not known.

4.8 Conclusion and Open Questions

In this chapter, we studied the FCFS $GI/GI/N$ queue in the Halfin-Whitt regime. We proved that under minor technical conditions the associated sequence of steady-state queue length distributions, normalized by $n^{\frac{1}{2}}$, is tight. We derived an upper bound for the large deviation exponent of the limiting steady-state queue length matching that conjectured in [43], and proved a matching lower bound for the case of Poisson arrivals. We also derived the first non-trivial bounds for the stochastic process studied in [91].

Our main proof technique was the derivation of new and simple bounds for the FCFS $GI/GI/n$ queue. Our bounds are of a structural nature, hold for all n and all times $t \geq 0$, and have intuitive closed-form representations as the suprema of certain natural processes which converge weakly to Gaussian processes. Our upper and lower bounds also exhibit a certain duality relationship, and exemplify a general methodology which may be useful for analyzing a variety of queueing systems.

This work leaves many interesting directions for future research. One pressing question is whether or not $\{n^{-\frac{1}{2}}\big(Q^n(\infty) - n\big)^+, n \geq 1\}$ has a *unique* weak limit. Similarly, although Corollary 21 shows that the process $\hat{Q}_1(t)$ is distributionally bounded over time, it is unknown whether $\hat{Q}_1(t)$ has a well-defined steady-state distribution. Furthermore, should $\{n^{-\frac{1}{2}}\big(Q^n(\infty) - n\big)^+, n \geq 1\}$ have a unique weak limit and $\hat{Q}_1(t)$ have a well-defined steady-state, must the two coincide? We note that similar questions (on the order of fluid, as opposed to diffusion, scaling) were investigated in [61], where the authors also handle systems with abandonments.

It is an open challenge to extend our techniques to more general models. For example, it would be interesting to generalize our lower bounds to non-Poisson arrival processes, as was done in [43] for the special case of processing times with finite support. It would also be interesting to generalize our bounds to systems with abandonments ($GI/GI/n + GI$). This setting is practically important, as the main application of the H-W regime has been to the study of call-centers, for which customer abandonments are an important modeling component [3]. For some interesting steps along these lines the reader is referred to the recent paper [28].

4.9 Appendix

4.9.1 Proof of Theorem 15

It is proven in [100] Theorem 1 (given in the notation of [100]) that

Theorem 22. *Suppose that for all sufficiently large n, $\{\zeta_{n,i}, i \geq 1\}$ is a stationary, countably infinite sequence of r.v. Let $a_n \overset{\Delta}{=} \mathbb{E}\big[\zeta_{n,1}\big]$, and $W_{n,k} \overset{\Delta}{=} \sum_{i=1}^{k} \zeta_{n,i}$. Further assume that $a_n < 0, \lim_{n\rightarrow\infty} a_n = 0$, and there exist $C_1, C_2 < \infty$ and $\epsilon > 0$ s.t. for all sufficiently large n,*

(i) $\mathbb{E}\big[|W_{n,k} - ka_n|^{2+\epsilon}\big] \leq C_1 k^{1+\frac{\epsilon}{2}}$ *for all* $k \geq 1$;

(ii) $\mathbb{P}\big(\max_{i=1,\ldots,k}(W_{n,i} - ia_n) > x\big) \leq C_2 \mathbb{E}\big[|W_{n,k} - ka_n|^{2+\epsilon}\big] x^{-(2+\epsilon)}$ *for all* $k \geq 1$ *and* $x > 0$;

(iii) $\mathbb{P}(\lim_{k\rightarrow\infty} W_{n,k} = -\infty) = 1$.

Then $\{|a_n| \sup_{k\geq 0} W_{n,k}, n \geq 1\}$ *is tight.*

With Theorem 22 in hand, we now complete the proof of Theorem 15.

Proof of Theorem 15. The proof follows almost exactly as the proof of Theorem 22 given in [100], and we now explicitly comment on precisely where the proof must be changed superficially so as to carry through under the slightly different set of assumptions of Theorem 15. First off, nowhere in the proof of Theorem 22 given in

[100] is assumption (iii) of Theorem 22 used, and thus that assumption is extraneous and may be removed. The only other difference between the set of assumptions for Theorem 22 and the set of assumptions for Theorem 15 is that assumption (ii) of Theorem 22 is replaced by assumption (ii) of Theorem 15. We now show that Theorem 22 holds under this change in assumptions. As in [100], let $x(a_n, k) \triangleq \frac{x}{|a_n|} + 2^k|a_n|$. Then the only place where assumption (ii) of Theorem 22 is used is between Equations 5 and 6, where this assumption is required to demonstrate that

$$\mathbb{P}\big(W_{n,2^k} - 2^k a_n > \frac{1}{2}x(a_n, k)\big) + \mathbb{P}\big(\max_{i=0,\ldots,2^k}\big(\sum_{j=1}^{i} \zeta_{n,j+2^k} - ia_n\big) > \frac{1}{2}x(a_n, k)\big) \tag{4.39}$$

$$\leq (1 + C_2)C_1 2^{2+\epsilon} 2^{k(1+\frac{\epsilon}{2})}\big(x(a_n, k)\big)^{-(2+\epsilon)}. \tag{4.40}$$

We now prove that assumption (ii) of Theorem 15 is sufficient to derive (4.40). In particular, the first summand of (4.39) is at most

$$\mathbb{E}\big[|W_{n,2^k} - 2^k a_n|^{2+\epsilon}\big]\big(\frac{1}{2}x(a_n, k)\big)^{-(2+\epsilon)} \text{ by Markov's inequality}$$

$$\leq C_1 2^{2+\epsilon} 2^{k(1+\frac{\epsilon}{2})}\big(x(a_n, k)\big)^{-(2+\epsilon)} \text{ by assumption (i) of Theorem 15.} \tag{4.41}$$

By the stationarity of $\{\zeta_{n,i}, i \geq 1\}$, the second summand of (4.39) equals

$$\mathbb{P}\big(\max_{i=0,\ldots,2^k}(W_{n,i} - ia_n) > \frac{1}{2}x(a_n, k)\big)$$

$$\leq C_2 2^{2+\epsilon} 2^{k(1+\frac{\epsilon}{2})}\big(x(a_n, k)\big)^{-(2+\epsilon)} \text{ by assumption (ii) of Theorem 15} \tag{4.42}$$

Since we may w.l.o.g. take $C_1, C_2 \geq 1$, it follows that $C_1 + C_2 \leq (1 + C_2)C_1$, and thus (4.40) follows from (4.41) and (4.42). The theorem follows from the proof of Theorem 22 given in [100]. □

4.9.2 Proof of Lemma 29

We note that the special case $r = 2$ is treated in [110]. Before proceeding with the proof of Lemma 29, it will be useful to prove three auxiliary results. The first treats the special case $n = 1, t \geq 1$ for ordinary (as opposed to equilibrium) renewal processes, and is proven in Theorem 1 of [21].

Theorem 23. *Suppose $Z(t)$ is an ordinary renewal process with renewal distribution X s.t. $\mathbb{E}[X] = \mu^{-1} \in (0,\infty)$, and $\mathbb{E}[X^r] < \infty$ for some $r \geq 2$. Then $\sup_{t\geq 1} t^{-\frac{r}{2}}\mathbb{E}\big[|Z(t) - \mu t|^r\big] < \infty$.*

Second, we prove a lemma treating the special case $n = 1, t \geq 1$ for equilibrium renewal processes.

Lemma 32. *Under the same definitions and assumptions as Lemma 29, for each $r \geq 2$, there exists $C_{X,r} < \infty$ (depending only on X and r) s.t. for all $t \geq 1$, $\mathbb{E}\big[|Z^e_1(t) - \mu t|^r\big] < C_{X,r}t^{\frac{r}{2}}$.*

Proof. Let X^e denote the first renewal interval in $Z^e_1(t)$, and f_{X^e} its density function, whose existence is guaranteed by (4.1). Observe that we may construct $Z^e_1(t)$ and an ordinary renewal process $Z(t)$ (also with renewal distribution X) on the same probability space so that for all $t \geq 0$, $Z^e_1(t) = I(X^e \leq t) + Z\big((t - X^e)^+\big)$, with $Z(t)$ independent of X^e. Thus

$$Z^e_1(t) - \mu t = \Big(Z\big((t - X^e)^+\big) - \mu\big(t - X^e\big)^+\Big) + \Big(I(X^e \leq t) - \mu\big(t - (t - X^e)^+\big)\Big).$$

Fixing some $t \geq 1$, it follows that $\mathbb{E}\big[|Z^e_1(t) - \mu t|^r\big]$ is at most

$$2^{r-1}\mathbb{E}\big[|Z\big((t - X^e)^+\big) - \mu\big(t - X^e\big)^+|^r\big] \tag{4.43}$$

$$+2^{r-1}\mathbb{E}\big[|I(X^e \leq t) - \mu\big(t - (t - X^e)^+\big)|^r\big] \text{ by the tri. ineq. and (4.8)} \tag{4.44}$$

We now bound the term $\mathbb{E}\big[|Z\big((t-X^e)^+\big)-\mu\big(t-X^e\big)^+|^r\big]$ appearing in (4.43), which equals

$$\int_0^{t-1}\mathbb{E}\big[|Z\big(t-s\big)-\mu\big(t-s\big)|^r\big]f_{X^e}(s)ds+\int_{t-1}^{t}\mathbb{E}\big[|Z\big(t-s\big)-\mu\big(t-s\big)|^r\big]f_{X^e}(s)ds. \quad (4.45)$$

Let $C'_{X,r}\overset{\Delta}{=}\sup_{t\geq 1}t^{-\frac{r}{2}}\mathbb{E}\big[|Z(t)-\mu t|^r\big]$. Theorem 23 implies that the first summand of (4.45) is at most

$$\begin{aligned}\int_0^{t-1}\big(C'_{X,r}(t-s)^{\frac{r}{2}}\big)f_{X^e}(s)ds &\leq \int_0^{t-1}\big(C'_{X,r}t^{\frac{r}{2}}\big)f_{X^e}(s)ds \\ &= C'_{X,r}t^{\frac{r}{2}}\mathbb{P}\big(X^e\leq t-1\big).\end{aligned}$$

Since $t-s\leq 1$ implies $|Z(t-s)-\mu(t-s)|^r\leq|Z(1)+\mu|^r$, the second summand of (4.45) is at most $\mathbb{E}\big[|Z(1)+\mu|^r\big]\mathbb{P}\big(X^e\in[t-1,t]\big)$. Combining our bounds for (4.45), we find that (4.43) is at most

$$2^{r-1}\mathbb{E}\big[|Z(1)+\mu|^r\big]+2^{r-1}C'_{X,r}t^{\frac{r}{2}}. \quad (4.46)$$

We now bound (4.44), which is at most

$$\begin{aligned}&2^{2r-2}\Big(1+\mathbb{E}\big[|\mu\big(t-(t-X^e)^+\big)|^r\big]\Big)\text{ by (4.8)} \\ &=2^{2r-2}\Big(1+\mu^r\big(\int_0^t s^r f_{X^e}(s)ds+\int_t^\infty t^r f_{X^e}(s)ds\big)\Big). \quad (4.47)\end{aligned}$$

It follows from (4.1) and Markov's inequality that for all $s\geq 0$, $f_{X^e}(s)=\mu\mathbb{P}(X>s)\leq\mu\mathbb{E}[X^r]s^{-r}$. Thus the term $\int_0^t s^r f_{X^e}(s)ds+\int_t^\infty t^r f_{X^e}(s)ds$ appearing in (4.47) is at most

$$\begin{aligned}\int_0^t s^r\big(\mu\mathbb{E}[X^r]s^{-r}\big)ds + t^r\int_t^\infty \big(\mu\mathbb{E}[X^r]s^{-r}\big)ds &= \mu\mathbb{E}[X^r]\Big(\int_0^t ds + t^r\int_t^\infty s^{-r}ds\Big)\\ &= \mu\mathbb{E}[X^r]\Big(t + t^r(r-1)^{-1}t^{1-r}\Big)\\ &= \mu\mathbb{E}[X^r]\big(1+(r-1)^{-1}\big)t. \qquad (4.48)\end{aligned}$$

Using (4.46) to bound (4.43) and (4.48) to bound (4.47) and (4.44), we find that $\mathbb{E}\big[|Z_1^e(t)-\mu t|^r\big]$ is at most

$$2^{r-1}\mathbb{E}\big[|Z(1)+\mu|^r\big] + 2^{r-1}C'_{X,r}t^{\frac{r}{2}} + 2^{2r-2} + 2^{2r-2}\mu^{r+1}\mathbb{E}[X^r]\big(1+(r-1)^{-1}\big)t. \quad (4.49)$$

Noting that $\mathbb{E}\big[|Z(1)+\mu|^r\big] < \infty$ since any renewal process, evaluated at any fixed time, has finite moments of all orders (see [86] p. 155), $\mathbb{E}[X^r] < \infty$ by assumption, and $t \leq t^{\frac{r}{2}}$ since $t \geq 1$ and $\frac{r}{2} \geq 1$, the lemma follows from (4.49). □

Third, we prove a lemma which will be useful in handling the case $t \leq 2$. We note that in this auxiliary lemma, the upper bound is of the form $(nt)^r$, as opposed to $(nt)^{\frac{r}{2}}$.

Lemma 33. *Under the same definitions and assumptions as Lemma 29, there exists $C_{X,r} < \infty$ (depending only on X and r) s.t. for all $n \geq 1$, and $t \in \big[0,2\big]$,*

$$\mathbb{E}\big[\big|\sum_{i=1}^n Z_i^e(t) - \mu nt\big|^r\big] \leq C_{X,r}\big(1+(nt)^r\big). \quad (4.50)$$

Proof. Note that the l.h.s. of (4.50) is at most

$$\mathbb{E}\big[\big|\sum_{i=1}^n Z_i^e(t) + \mu nt\big|^r\big] \;\leq\; 2^{r-1}\Big(\mathbb{E}\big[\big(\sum_{i=1}^n Z_i^e(t)\big)^r\big] + (\mu nt)^r\Big) \text{ by (4.8)}. \quad (4.51)$$

We now bound the term $\mathbb{E}\big[\big(\sum_{i=1}^{n} Z_i^e(t)\big)^r\big]$ appearing in (4.51). Let $\{Z_i(t)\}$ denote a countably infinite sequence of i.i.d. ordinary renewal processes with renewal distribution X. Let us fix some $t \in [0,2]$ and $n \geq 1$, and let $\{B_i\}$ denote a countably infinite sequence of i.i.d. Bernoulli r.v. s.t $\mathbb{P}(B_i = 1) = p \triangleq \mathbb{P}(R(X) \leq t)$. Note that we may construct $\{Z_i^e(t)\}$, $\{Z_i(t)\}, \{B_i\}$ on the same probability space so that w.p.1 $Z_i^e(t) \leq B_i\big(1 + Z_i(t)\big)$ for all $i \geq 1$, with $\{Z_i(t)\}, \{B_i\}$ mutually independent. Letting $M \triangleq \sum_{i=1}^{n} B_i$, it follows that

$$\mathbb{E}[(\sum_{i=1}^{n} Z_i^e(t))^{\lceil r \rceil}] \leq \mathbb{E}\Big[\Big(\sum_{i=1}^{M} (1 + Z_i(t))\Big)^{\lceil r \rceil}\Big]. \tag{4.52}$$

Let Z^+ denote the set of non-negative integers. Note that for any positive integer k,

$$\begin{aligned} \mathbb{E}\Big[\Big(\sum_{i=1}^{k} (1 + Z_i(t))\Big)^{\lceil r \rceil}\Big] &= \mathbb{E}\Big[\sum_{\substack{j_1,\ldots,j_k \in Z^+ \\ j_1+\ldots+j_k = \lceil r \rceil}} \prod_{i=1}^{k} (1 + Z_i(t))^{j_i}\Big] \\ &= \sum_{\substack{j_1,\ldots,j_k \in Z^+ \\ j_1+\ldots+j_k = \lceil r \rceil}} \prod_{i=1}^{k} \mathbb{E}\big[(1 + Z_i(t))^{j_i}\big] \end{aligned} \tag{4.53}$$

For any setting of $\{j_i, i = 1, \ldots, k\}$ in the r.h.s. of (4.53), at most $\lceil r \rceil$ of the j_i are strictly positive, and each j_i is at most $\lceil r \rceil$. It follows that the term $\prod_{i=1}^{k} \mathbb{E}\big[(1 + Z_i(t))^{j_i}\big]$ appearing in the r.h.s. of (4.53) is at most $\Big(\mathbb{E}\big[(1+Z_1(t))^{\lceil r \rceil}\big]\Big)^{\lceil r \rceil}$, irregardless of the particular setting of $\{j_i, i = 1, \ldots, k\}$. As there are a total of $k^{\lceil r \rceil}$ distinct feasible configurations for $\{j_i, i = 1, \ldots, k\}$ in the r.h.s. of (4.53), combining the above we find that for any non-negative integer k,

$$\begin{aligned}\mathbb{E}\Big[\Big(\sum_{i=1}^{k}(1+Z_i(t))\Big)^{\lceil r\rceil}\Big] &\leq k^{\lceil r\rceil}\Big(\mathbb{E}\big[(1+Z_1(t))^{\lceil r\rceil}\big]\Big)^{\lceil r\rceil} \\ &\leq k^{\lceil r\rceil}\Big(\mathbb{E}\big[(1+Z_1(2))^{\lceil r\rceil}\big]\Big)^{\lceil r\rceil} \text{ since } t \leq 2. \qquad (4.54)\end{aligned}$$

Since any renewal process, evaluated at any fixed time, has finite moments of all orders (see [86] p. 155), it follows that $C^1_{X,\lceil r\rceil} \triangleq \Big(\mathbb{E}\big[(1+Z_1(2))^{\lceil r\rceil}\big]\Big)^{\lceil r\rceil}$ is a finite constant depending only on X and $\lceil r\rceil$. Combining (4.52) and (4.54) with the independence of M and $\{Z_i(t)\}$, it follows from a simple conditioning argument that

$$\mathbb{E}\big[\big(\sum_{i=1}^{n} Z_i^e(t)\big)^{\lceil r\rceil}\big] \leq C^1_{X,\lceil r\rceil}\mathbb{E}\big[M^{\lceil r\rceil}\big]. \qquad (4.55)$$

We now bound the term $\mathbb{E}\big[M^{\lceil r\rceil}\big]$ appearing in (4.55). Noting that M is a binomial distribution with parameters n and p, it follows from [92] Equation 3.3 that there exist finite constants $C_{0,\lceil r\rceil}, C_{1,\lceil r\rceil}, C_{2,\lceil r\rceil}, \ldots, C_{\lceil r\rceil,\lceil r\rceil}$, independent of n and p, s.t. $\mathbb{E}\big[M^{\lceil r\rceil}\big] = \sum_{k=0}^{\lceil r\rceil} C_{k,\lceil r\rceil} p^k \prod_{j=0}^{k-1}(n-j)$. Further noting that $\prod_{j=0}^{k-1}(n-j) \leq n^k$ for all $k \geq 0$, it follows that $\mathbb{E}\big[M^{\lceil r\rceil}\big] \leq \sum_{k=0}^{\lceil r\rceil} |C_{k,\lceil r\rceil}|(np)^k$. Letting $C^2_{\lceil r\rceil} \triangleq \max_{i=0,\ldots,\lceil r\rceil} |C_{i,\lceil r\rceil}|$, it follows from (4.55) that

$$\begin{aligned}\mathbb{E}\big[\big(\sum_{i=1}^{n} Z_i^e(t)\big)^{\lceil r\rceil}\big] &\leq C^1_{X,\lceil r\rceil} C^2_{\lceil r\rceil} \sum_{i=0}^{\lceil r\rceil}(np)^i \\ &\leq C^1_{X,\lceil r\rceil} C^2_{\lceil r\rceil}(\lceil r\rceil + 1)(1+np)^{\lceil r\rceil}. \qquad (4.56)\end{aligned}$$

Recall that for any non-negative r.v. Y, one has that $\mathbb{E}[Y^r] \leq \mathbb{E}[Y^{\lceil r\rceil}]^{\frac{r}{\lceil r\rceil}}$. Thus letting

$C^3_{X,r} \triangleq \left(C^1_{X,\lceil r \rceil} C^2_{\lceil r \rceil} (\lceil r \rceil + 1) \right)^{\frac{r}{\lceil r \rceil}}$, it follows from (4.56) that

$$\mathbb{E}\big[\big(\sum_{i=1}^{n} Z_i^e(t)\big)^r\big] \leq C^3_{X,r}(1+np)^r. \tag{4.57}$$

Furthermore, it follows from (4.1) that $p = \mu \int_0^t \mathbb{P}(X > y)dy \leq \mu t$. Combining with (4.57), we find that

$$\mathbb{E}\big[\big(\sum_{i=1}^{n} Z_i^e(t)\big)^r\big] \leq C^3_{X,r}(1+\mu nt)^r \tag{4.58}$$

Plugging (4.58) back into (4.51), it follows that the l.h.s. of (4.50) is at most

$$2^{r-1}\left(C^3_{X,r}(1+\mu nt)^r + (\mu nt)^r \right) \leq 2^r (C^3_{X,r} + 1)(1+\mu nt)^r.$$

Noting that $(1+\mu nt)^r \leq 2^r\big(1+(\mu nt)^r\big)$ by (4.8), and $1+(\mu nt)^r \leq (1+\mu)^r\big(1+(nt)^r\big)$, completes the proof. □

With the above auxiliary results in hand, we now complete the proof of Lemma 29.

Proof of Lemma 29. We proceed by a case analysis. First, suppose $t \leq \frac{2}{n}$. Then we also have $t \leq 2$, and by Lemma 33 there exists $C^1_{X,r} < \infty$ s.t. the l.h.s. of (4.9) is at most

$$C^1_{X,r}\big(1+(nt)^r\big) \quad \leq \quad C^1_{X,r}(1+2^r) \qquad \text{since } t \leq \frac{2}{n} \text{ implies } nt \leq 2.$$

Letting $M_1 \triangleq C^1_{X,r}(1+2^r)$, it follows that the l.h.s. of (4.9) is at most $M_1 \leq M_1\big(1+(nt)^{\frac{r}{2}}\big)$, completing the proof for the case $t \leq \frac{2}{n}$.

Second, suppose $t \in \big[\frac{2}{n}, 2\big]$. Let $n_1(t) \triangleq \lfloor nt \rfloor$. Noting that $t \geq \frac{2}{n}$ implies $n_1(t) > 0$, in this case we may define $n_2(t) \triangleq \lfloor \frac{n}{n_1(t)} \rfloor$. Then the l.h.s. of (4.9) equals

$$\begin{aligned}
& \mathbb{E}[|\sum_{m=1}^{n_1(t)}\sum_{l=1}^{n_2(t)}\big(Z^e_{(m-1)n_2(t)+l}(t)-\mu t\big) + \sum_{l=n_1(t)n_2(t)+1}^{n}\big(Z^e_l(t)-\mu t\big)|^r] \\
\leq \quad & 2^{r-1}\mathbb{E}[|\sum_{m=1}^{n_1(t)}\sum_{l=1}^{n_2(t)}\big(Z^e_{(m-1)n_2(t)+l}(t)-\mu t\big)|^r] \qquad (4.59) \\
& +2^{r-1}\mathbb{E}[|\sum_{l=n_1(t)n_2(t)+1}^{n}\big(Z^e_l(t)-\mu t\big)|^r] \quad \text{by the tri. ineq. and (4.8)} \qquad (4.60)
\end{aligned}$$

We now bound (4.59). By Lemma 28, there exists $C_r < \infty$ s.t. (4.59) is at most

$$\begin{aligned}
& 2^{r-1}C_r(n_1(t))^{\frac{r}{2}}\mathbb{E}[|\sum_{l=1}^{n_2(t)}\big(Z^e_l(t)-\mu t\big)|^r] \\
\leq \quad & 2^{r-1}C_r(n_1(t))^{\frac{r}{2}}\Bigg(C^1_{X,r}\Big(1+\big(n_2(t)t\big)^r\Big)\Bigg) \quad \text{by Lemma 33, since } t \leq 2. \qquad (4.61)
\end{aligned}$$

We now bound the term $tn_2(t)$ appearing in (4.61). In particular,

$$\begin{aligned}
tn_2(t) &= t\lfloor \frac{n}{\lfloor nt \rfloor} \rfloor \\
&\leq \frac{nt}{\lfloor nt \rfloor} \leq \frac{nt}{nt-1}. \qquad (4.62)
\end{aligned}$$

But since $t \geq \frac{2}{n}$ imples $nt \geq 2$, and $g(z) \triangleq \frac{z}{z-1}$ is a decreasing function of z on $(1, \infty)$, it follows from (4.62) that

$$tn_2(t) \leq 2.$$

Since $n_1(t) \leq nt$, it thus follows from (4.61) that (4.59) is at most

$$2^{r-1}C_rC^1_{X,r}(1+2^r)(nt)^{\frac{r}{2}}. \qquad (4.63)$$

We now bound (4.60). Note that the sum $\sum_{l=n_1(t)n_2(t)+1}^{n}\big(Z_l^e(t)-\mu t\big)$ appearing in (4.60) is taken over $n-n_1(t)n_2(t)$ terms. Furthermore,

$$\begin{aligned} n-n_1(t)n_2(t) &= n-n_1(t)\lfloor\frac{n}{n_1(t)}\rfloor \\ &\leq n-n_1(t)\big(\frac{n}{n_1(t)}-1\big) \\ &= n_1(t). \end{aligned}$$

As $n_1(t)\leq nt$, it thus follows from Lemma 28 that (4.60) is at most

$$\begin{aligned} 2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[|Z_1^e(t)-\mu t|^r\big] &\leq 2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[\big(Z_1^e(t)+\mu t\big)^r\big] \\ &\leq 2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[\big(Z_1^e(2)+2\mu\big)^r\big] \text{ since } t\leq 2 \quad (4.64) \end{aligned}$$

Using (4.63) to bound (4.59) and (4.64) to bound (4.60) shows that the l.h.s. of (4.9) is at most

$$2^{r-1}C_rC_{X,r}^1(1+2^r)(nt)^{\frac{r}{2}}+2^{r-1}C_r(nt)^{\frac{r}{2}}\mathbb{E}\big[\big(Z_1^e(2)+2\mu\big)^r\big]. \qquad (4.65)$$

Let $M_2\triangleq 2^{r-1}C_rC_{X,r}^1(1+2^r)+2^{r-1}C_r\mathbb{E}\big[\big(Z_1^e(2)+2\mu\big)^r\big]$. It follows from (4.65) that the l.h.s. of (4.9) is at most $M_2(nt)^{\frac{r}{2}}\leq M_2\big(1+(nt)^{\frac{r}{2}}\big)$, completing the proof for the case $t\in\big[\frac{2}{n},2\big]$.

Finally, suppose $t\geq 2$. In this case, it follows from Lemma 28 that the l.h.s. of (4.9) is at most $C_rn^{\frac{r}{2}}\mathbb{E}\big[|Z_1^e(t)-\mu t|^r\big]$. Let $C_{X,r}^2\triangleq\sup_{t\geq 2}t^{-\frac{r}{2}}\mathbb{E}\big[|Z_1^e(t)-\mu t|^r\big]$. Then it follows from Lemma 32 that $C_{X,r}^2<\infty$, and the l.h.s. of (4.9) is at most $C_rC_{X,r}^2(nt)^{\frac{r}{2}}$. Letting $M_3\triangleq C_rC_{X,r}^2$, it follows that the l.h.s. of (4.9) is at most $M_3(nt)^{\frac{r}{2}}\leq M_3\big(1+(nt)^{\frac{r}{2}}\big)$, completing the proof for the case $t\geq 2$.

As this treats all cases, we can complete the proof of the lemma by letting $M_4 \triangleq \max\big(M_1, M_2, M_3\big)$, and noting that for all $n \geq 1$ and $t \geq 0$, the l.h.s. of (4.9) is at most $M_4\big(1 + (nt)^{\frac{r}{2}}\big)$. □

Chapter 5

Probability of Delay for the Steady-state GI/GI/n Queue in the Halfin-Whitt Regime

5.1 Introduction and Literature Review

Recall from Sections 1.3 and 4.1 that the H-W regime was formally introduced by Halfin and Whitt [52], who studied the $GI/M/n$ system (for large n) when the traffic intensity scales like $1 - Bn^{-\frac{1}{2}}$ for some strictly positive B. As described in [52],[45], an important motivation for the H-W regime are applications in which the system designer wishes some non-trivial fraction of all jobs to have to wait for service, where this fraction should be bounded away from both zero and unity even as the system increases in size. Furthermore, this steady-state probability of delay often appears in objective functions used to capture the quality-efficiency trade-off in the H-W regime [10],[76]. The steady-state probability of delay for exponentially distributed processing times was computed explicitly by Halfin and Whitt in [52], and an explicit formula is also known for the case of deterministic processing times [60]. Gamarnik and Momcilovic give an implicit description (in terms of a certain Markov chain) of the steady-state probability of delay for the case of processing times with finite support, and prove that this probability lies strictly in $(0, 1)$, in [43]. However, it seems that essentially nothing was known about this important quantity for more general processing time distributions prior to this thesis. Furthermore, other than for the special cases of deterministic [60] or exponentially distributed [52] processing times, it seems that little was known about the qualitative behavior and scaling of this probability, beyond its lying strictly in $(0, 1)$ for the case of processing times with finite support [43].

Another question of interest in the analysis of parallel server queues in the H-W regime is large deviations behavior, i.e. the probability of certain rare events. Recall that we proved several results along these lines in Chapter 4, and that this quantity had been studied previously be several authors [43],[52],[60]. However, other than for the special case of exponentially distributed [52] and deterministic [60] processing times, all of these results had been about the probability of seeing an exceptionally large number in system, as opposed to seeing an exceptionally small number in system, i.e. an exceptionally large number of idle servers.

In this chapter we prove the first qualitative results about the steady-state probability of delay for generally distributed processing times. In particular, under very minor technical conditions, we derive bounds on this probability as $B \rightarrow \infty$ and $B \rightarrow 0$ for fixed inter-arrival and processing time distributions. As we will see, it follows from known results for the case of exponentially distributed inter-arrival or processing times that our bounds are, in a sense, tight. We also revisit the question of large deviations for the steady-state $GI/GI/n$ queue in the H-W regime, but now examine the probability that the steady-state number of idle servers exceeds some large value x, deriving bounds on this probability as $x \rightarrow \infty$, which are again tight in an appropriate sense.

As in Chapter 4, our main proof technique is the derivation of new and simple bounds for the FCFS $GI/GI/n$ queue. Our bounds are of a structural nature, hold for all n and all times $t \geq 0$, and have intuitive closed-form representations as the suprema of certain natural processes which converge weakly to Gaussian processes.

5.1.1 Outline of chapter

The rest of the chapter proceeds as follows. In Section 5.2, we present our main results. In Section 5.3, we establish our general-purpose upper bounds for the queue length in a properly initialized FCFS $GI/GI/n$ queue. In Section 5.4, we prove an asymptotic version of our upper bound in the H-W regime. In Section 5.5, we prove our bounds on the steady-state probability of delay as $B \rightarrow \infty$. In Section 5.6, we prove our bounds on the steady-state probability of delay as $B \rightarrow 0$. In Section 5.7, we prove our bounds on the large deviations behavior of the steady-state number of idle servers. In Section 5.8, we compare to previous results from the literature, which show that our bounds are tight in an appropriate sense. In Section 5.9 we summarize our main results and comment on directions for future research. We include a technical appendix in Section 5.10.

5.2 Main Results

As in Chapter 4, $\lambda_{n,B} = n - Bn^{\frac{1}{2}}$, and Q^n_B is the First-Come-First-Serve (FCFS) $GI/GI/n$ queue with inter-arrival times drawn i.i.d. distributed as $A\lambda^{-1}_{n,B}$ and processing times drawn i.i.d. distributed as S (initial conditions will be specified later). Suppose that $\mathbb{E}[A] = \mu_A^{-1} < \infty, \mathbb{E}[S] = \mu_S^{-1} < \infty$, and $\mathbb{P}(A = 0) = \mathbb{P}(S = 0) = 0$. Recall that σ^2_A and σ^2_S denote the variances of A and S, respectively, and c^2_A and

c_S^2 denote the squared coefficients of variation (s.c.v.) of A and S, respectively. All processes should be assumed right-continuous with left limits (r.c.l.l.) unless stated otherwise. All empty summations should be evaluated as zero, and all empty products should be evaluated as one.

Recall that in Section 4.2 we defined two sets of assumptions, the H-W assumptions and the T_0 assumptions, for A and S. The H-W assumptions ensured that $\{Q_B^n, n \geq 1\}$ was in the H-W regime. The T_0 assumptions were a set of additional minor technical conditions. In this chapter, we will again refer to these assumptions, and refer the reader to Section 4.2 for details. For clarity of exposition, statements of (in)equality w.p.1 are not distinguished from statements of (in)equality.

5.2.1 Main results

We now state our main results. We begin by stating our bound on the steady-state probability of delay as $B \to \infty$.

Theorem 24. *For any fixed A and S which satisfy the $H-W$ and T_0 assumptions,*

$$\limsup_{B\to\infty} B^{-2} \log\left(\limsup_{n\to\infty} \mathbb{P}\big(Q_B^n(\infty) \geq n\big)\right) < 0.$$

In words, Theorem 24 states that there exists $\epsilon > 0$, depending only on A and S, s.t. the limiting steady-state probability of delay is bounded from above by $\exp\big(-\epsilon_1 B^2\big)$ as $B \to \infty$. We now give our bounds on the steady-state probability of delay as $B \to 0$.

Theorem 25. *For any fixed A and S which satisfy the $H-W$ and T_0 assumptions, s.t. in addition $\mathbb{E}[S^3] < \infty$ and $c_A^2 > 0$,*

$$\liminf_{B\to 0} B^{-1} \liminf_{n\to\infty} \mathbb{P}\big(Q_B^n(\infty) < n\big) > 0.$$

In words, Theorem 24 states that there exists $\epsilon > 0$, depending only on A and S, s.t. the limiting steady-state probability that a job does not have to wait for service, i.e. no delay, is bounded from below by ϵB as $B \rightarrow 0$. Equivalently, the limiting steady-state probability of delay is bounded from above by $1 - \epsilon B$ as $B \rightarrow 0$. We note that this result is somewhat surprising, since in light of Theorem 24, one might expect this probability to scale as B^2 as $B \rightarrow 0$, since $1 - \exp(-\epsilon B^2)$ behaves like ϵB^2 as $B \rightarrow 0$.

We now state our bounds on the large deviations behavior for the number of idle servers.

Theorem 26. *For any fixed A and S which satisfy the $H - W$ and T_0 assumptions, s.t. in addition $c_A^2 > 0$, and any fixed $B > 0$,*

$$\liminf_{x \rightarrow \infty} x^{-2} \log \left(\liminf_{n \rightarrow \infty} \mathbb{P}\Big(\big(Q^n(\infty) - n\big) n^{-\frac{1}{2}} < -x \Big) \right) > -\infty.$$

In words, Theorem 26 states that there exists $\epsilon > 0$, depending only on A, S, and B, s.t. the tail of the limiting steady-state number of idle servers is bounded from below by $\exp\big(- \epsilon x^2\big)$ as $x \rightarrow \infty$.

In Section 5.8, we will show that our results give the correct scaling for the case of exponentially distributed inter-arrival or processing times. We note that our results can not be derived using naive infinite-server lower bounds, as in all cases our inequalities point in the other direction. Also, as in Chapter 4, our results translate into bounds for any weak limit of the sequence $\{\big(Q^n(\infty) - n\big)^+ n^{-\frac{1}{2}}, n \geq 1\}$.

5.3 Upper Bound

In this section, we prove general upper bounds for the FCFS $GI/GI/n$ queue, when properly initialized. The bounds are valid for all finite n, and work in both the transient and steady-state (when it exists) regimes. Although we will later customize these bounds to the H-W regime to prove our main results, we note that the bounds are in no way limited to that regime. Recall that for a non-negative r.v. X with finite mean $\mathbb{E}[X] > 0$, $R(X)$ denotes a r.v. distributed as the residual life distribution of X; see (4.1) for details. Recall that associated with a non-negative r.v. X, an equilibrium renewal process with renewal distribution X is a counting process in which the first inter-event time is distributed as $R(X)$, and all subsequent inter-event times are drawn i.i.d. distributed as X; an ordinary renewal process with renewal distribution X is a counting process in which all inter-event times are drawn i.i.d. distributed as X. As in Chapter 4, let $\{N_i(t), i = 1, \dots, n\}$ denote a set of n i.i.d. equilibrium renewal processes with renewal distribution S. Let $A(t)$ denote an independent equilibrium renewal process with renewal distribution A. For $s \in \mathbb{R}^+$, let $V_i^1(s)$ denote the remaining time (at time s) until the first renewal to occur after time s in process $N_i(t)$, $i = 1, \dots, n$. Let $V_i^j(s)$ denote the length of the $\big(N_i(s) + j\big)$th renewal interval in process $N_i(t)$, $j \geq 2$, $i = 1, \dots, n$. Namely, $V_i^j(s)$ is the length of the $(j-1)$th renewal interval to be initiated in process $N_i(t)$ after time s. Similarly, let $U^1(s)$ denote the remaining time (at time s) until the first renewal to occur after time s in process $A(t)$, and $U^j(s)$ denote the length of the $\big(A(s) + j\big)$th renewal interval in process $A(t)$, $j \geq 2$. For $x \in \mathbb{R}^+$, let $A^x(t) \overset{\Delta}{=} A(x, x+t)$, $dA^x(t) \overset{\Delta}{=} A^x(t) - A^x(t^-)$, and $A^x(s,t) \overset{\Delta}{=} A^x(t) - A^x(s)$. Let $N_i^x(t) \overset{\Delta}{=} N_i(x, x+t)$, $dN_i^x(t) \overset{\Delta}{=} N_i^x(t) - N_i^x(t^-)$, $N_i^x(s,t) \overset{\Delta}{=} N_i^x(t) - N_i^x(s)$, $i = 1, \dots, n$. For $z \in Z^+$ s.t. $z \leq n$, let $\tau_{z,0}^x \overset{\Delta}{=} 0$, and let $\{\tau_{z,k}^x, k \geq 1\}$ denote the sequence of event times in the pooled renewal process $A^x(t) + \sum_{i=1}^z N_i^x(t)$. For $y \in \mathbb{R}^+$, let

$$\phi(x, y, z) \overset{\Delta}{=} \sup_{0 \leq s \leq y} \big(A^x(y-s, y) - \sum_{i=1}^{z} N_i^x(y-s, y)\big).$$

Let $V_i^j \triangleq V_i^j(0)$, and $U^j \triangleq U^j(0)$, $i = 1, \ldots, n$, $j \geq 1$.

For $v \in \mathbb{R}^+$, and $\eta \in Z^+$ s.t. $\eta \leq n$, let $\mathcal{Q}_\eta^v$ denote the FCFS $GI/GI/\eta$ queue with inter-arrival times drawn i.i.d. distributed as A, processing times drawn i.i.d. distributed as S, and the following initial conditions. For $i = 1, \ldots, \eta$, there is a single job initially being processed on server i, with initial processing time $V_i^1(v)$. There are $\phi(0, v, n)$ jobs waiting in queue, and the first inter-arrival time is $U^1(v)$. We let $Q_\eta^v(t)$ denote the number in system in $\mathcal{Q}_\eta^v$ at time t. We also let $\mathcal{Q} \triangleq \mathcal{Q}_n^0$, and $Q(t) \triangleq Q_n^0(t)$. We now establish an upper bound for $Q(t)$.

Theorem 27. *For all $t, x \geq 0$, $\mathbb{P}\big(Q(t) > x\big)$ is at most*

$$\inf_{\substack{\delta \in [0,t] \\ \eta \in [0,n]}} \mathbb{P}\Bigg(\max\bigg(1 + \sup_{0 \leq s \leq \delta} \big(A(s) - \sum_{i=1}^{\eta} N_i(s)\big), \\ \sup_{\delta \leq s \leq t} \big(A(s) - \sum_{i=1}^{n} N_i(s)\big) + \sum_{i=\eta+1}^{n} N_i(\delta)\bigg) \\ + \sum_{i=\eta+1}^{n} I(N_i(\delta) = 0) \quad > \quad x - \eta \Bigg).$$

If in addition $Q(t)$ converges weakly to a steady-state distribution $Q(\infty)$ as $t \to \infty$, then for all $x > 0$, $\mathbb{P}\big(Q(\infty) > x\big)$ is at most

$$\inf_{\substack{\delta \geq 0 \\ \eta \in [0,n]}} \mathbb{P}\Bigg(\max\bigg(1 + \sup_{0 \leq t \leq \delta} \big(A(t) - \sum_{i=1}^{\eta} N_i(t)\big), \\ \sup_{t \geq \delta} \big(A(t) - \sum_{i=1}^{n} N_i(t)\big) + \sum_{i=\eta+1}^{n} N_i(\delta)\bigg) \\ + \sum_{i=\eta+1}^{n} I(N_i(\delta) = 0) \quad > \quad x - \eta \Bigg).$$

Recall that in Chapter 4 we examined a modified queueing system in which all servers were kept busy at all times by adding artificial arrivals whenever a server would otherwise go idle. We will prove Theorem 27 by analyzing a different modified queueing system, in which all servers are kept busy on some fixed time interval $[0, t-\delta]$ by adding artificial arrivals, at the end of that time interval servers $\eta+1, \ldots, n$ break down and cease functioning, and for the remaining time the remaining functional servers $1, \ldots, \eta$ are again kept busy by adding artificial arrivals. By simultaneously altering the number of servers and keeping all servers busy, we will be able to derive non-trivial bounds for both the steady-state probability of delay, and the probability of there being many idle servers. We note that the upper bound of Chapter 4, namely Theorem 13, can be recovered as a special case of Theorem 27, modulo an additional "$+\,1$" which appears in the statement of Theorem 27. Indeed, by setting $\delta = 0$ and $\eta = n$, no servers break down during the time horizon $[0, t]$, and all n servers are kept busy on the entire time horizon $[0, t]$, which is exactly the bounding system of Chapter 4.

We begin by defining two auxiliary processes $\tilde{A}^{v}_{\gamma,\eta}(t)$ and $\tilde{Q}^{v}_{\gamma,\eta}(t)$, where $\tilde{A}^{v}_{\gamma,\eta}(t)$ will become the arrival process to $\tilde{\mathcal{Q}}^{v}_{\gamma,\eta}$, and we will later prove that $\tilde{Q}^{v}_{\gamma,\eta}(t)$ equals the number in system in $\tilde{\mathcal{Q}}^{v}_{\gamma,\eta}$ at time t. Whenever there is no ambiguity, we use the notations $\tau_k, \tilde{\mathcal{Q}}, \tilde{A}(t)$, and $\tilde{Q}(t)$ as shorthand for $\tau^{v}_{\eta,k}$, $\tilde{\mathcal{Q}}^{v}_{\gamma,\eta}$, $\tilde{A}^{v}_{\gamma,\eta}(t)$, and $\tilde{Q}^{v}_{\gamma,\eta}(t)$ respectively. We note that to prove Theorem 27, we must allow our servers to be initialized at a general time v in the corresponding renewal processes, to capture the dependencies between the remaining processing times of the jobs in service at time γ (equivalently $t-\delta$) and the number of jobs waiting in queue at time γ.

We now define the processes $\tilde{A}(t)$ and $\tilde{Q}(t)$ on $[0, \gamma]$ inductively over $\{\tau_k, k \geq 0\}$. Let $\tilde{A}(\tau_0) \triangleq 0$, $\tilde{Q}(\tau_0) \triangleq \eta + \phi(0, v, n)$. Now suppose that for some $k \geq 0$, we have defined $\tilde{A}(t)$ and $\tilde{Q}(t)$ for all $t \leq \tau_k$, and $\tau_{k+1} \leq \gamma$. We now define these processes for $t \in (\tau_k, \tau_{k+1}]$. For $t \in (\tau_k, \tau_{k+1})$, let $\tilde{A}(t) \triangleq \tilde{A}(\tau_k)$, and $\tilde{Q}(t) \triangleq \tilde{Q}(\tau_k)$. Note that $dA^{v}(\tau_{k+1}) + \sum_{i=1}^{\eta} dN^{v}_{i}(\tau_{k+1}) = 1$, since $R(X)$ and $R(A)$ are continuous r.v.s,

$\mathbb{P}(A = 0) = \mathbb{P}(S = 0) = 0$, and $A(t), \{N_i(t), i = 1, \ldots, \eta\}$ are mutually independent and have stationary increments. We define

$$\tilde{A}(\tau_{k+1}) \triangleq \begin{cases} \tilde{A}(\tau_k) + 1 & \text{if } dA^v(\tau_{k+1}) = 1; \\ \tilde{A}(\tau_k) + 1 & \text{if } \sum_{i=1}^{\eta} dN_i^v(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) \leq \eta; \\ \tilde{A}(\tau_k) & \text{otherwise (i.e. } \sum_{i=1}^{\eta} dN_i^v(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) > \eta); \end{cases}$$

Similarly, we define

$$\tilde{Q}(\tau_{k+1}) \triangleq \begin{cases} \tilde{Q}(\tau_k) + 1 & \text{if } dA^v(\tau_{k+1}) = 1; \\ \tilde{Q}(\tau_k) & \text{if } \sum_{i=1}^{\eta} dN_i^v(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) \leq \eta; \\ \tilde{Q}(\tau_k) - 1 & \text{otherwise (i.e. } \sum_{i=1}^{\eta} dN_i^v(\tau_{k+1}) = 1 \text{ and } \tilde{Q}(\tau_k) > \eta); \end{cases}$$

Also, for $k = A^v(\gamma) + \sum_{i=1}^{z} N_i^v(\gamma)$, namely the largest index s.t. $\tau_k \leq \gamma$, we let $\tilde{A}(t) \triangleq \tilde{A}(\tau_k)$ for $t \in (\tau_k, \gamma]$, and $\tilde{Q}(t) \triangleq \tilde{Q}(\tau_k)$ for $t \in (\tau_k, \gamma]$. Combining the above completes our inductive definition of $\tilde{A}(t)$ and $\tilde{Q}(t)$ on $[0, \gamma]$, since $\lim_{k \to \infty} \tau_k = \infty$. We now define $\tilde{A}(t)$ on (γ, ∞). For all $t > \gamma$, $d\tilde{A}(t) = dA^v(t)$, namely the events in the two processes coincide. It follows from our construction that both $\tilde{A}(t)$ and $\tilde{Q}(t)$ are well-defined and r.c.l.l. on $[0, \gamma)$, and $\tilde{A}(t)$ is well-defined and r.c.l.l. on $[0, \infty)$.

We now construct the FCFS $G/GI/\eta$ queue $\tilde{\mathcal{Q}}$ using the auxiliary process $\tilde{A}(t)$, and define $\tilde{Q}(t)$ on $[\gamma, \infty)$. $\tilde{\mathcal{Q}}$ is defined to be the FCFS $G/GI/\eta$ queue with arrival process $\tilde{A}(t)$ and processing time distribution S, where the jth job assigned to server i (after time 0) is assigned processing time $V_i^{j+1}(v)$ for $j \geq 1, i = 1, \ldots, \eta$, and jobs are always assigned to the available server of least index. The initial conditions for $\tilde{\mathcal{Q}}$ are s.t. for $i = 1, \ldots, \eta$, there is a single job initially being processed on server i with initial processing time $V_i^1(v)$, and there are $\phi(0, v, n)$ jobs waiting in queue. For $t \geq \gamma$, we define $\tilde{Q}(t)$ to be the number in system in $\tilde{\mathcal{Q}}$ at time t. Note that on

$[\gamma,\infty)$, $\tilde{Q}$ operates like a 'normal' FCFS $GI/GI/\eta$ queue.

We now analyze $\tilde{Q}$. The system is nearly identical to the bounding system considered in Chapter 4, the only change being that here we allow for more general initial conditions, and restrict our analysis to the interval $[0,\gamma]$. The following lemma is essentially identical to Chapter 4, Lemma 25. The proof follows nearly identically to the proof of Lemma 25, and we refer the reader to Chapter 4, Section 4.3, Lemma 25 for details.

Lemma 34. *For $i=1,\ldots,\eta$, exactly one job departs from server i at each time $t \in \{\sum_{l=1}^{j} V_i^l(v), j \geq 1\} \bigcap [0,\gamma]$, and there are no other departures from server i on $[0,\gamma]$. Also, no server ever idles in $\tilde{Q}$ on $[0,\gamma]$, $\tilde{Q}(t)$ equals the number in system in $\tilde{Q}$ at time t for all $t \leq \gamma$, and for all k s.t. $\tau_k \leq \gamma$,*

$$\tilde{Q}(\tau_k) - \eta = \max\Big(0, \tilde{Q}(\tau_{k-1}) - \eta + dA^v(\tau_k) - \sum_{i=1}^{\eta} dN_i^v(\tau_k)\Big). \tag{5.1}$$

Note that it follows from Lemma 34 and our definition of $\tilde{Q}(t)$ on $[\gamma,\infty)$ that $\tilde{Q}(t)$ is r.c.l.l. on $[0,\infty)$, and

Corollary 6. *$\tilde{Q}(t)$ equals the number in system in $\tilde{Q}$ at time t for all $t \geq 0$.*

We now 'unfold' recursion (5.1) to derive a simple one-dimensional random walk representation for $\tilde{Q}(t)$, $t \leq \gamma$. It follows from (5.1) and a straightforward induction on $\{\tau_k, k \geq 0\}$ that for all k s.t. $\tau_k \leq \gamma$, $\tilde{Q}(\tau_k) - \eta$ equals

$$\max\Big(\max_{j\in[0,k-1]}\big(A^v(\tau_{k-j},\tau_k) - \sum_{i=1}^{\eta} N_i^v(\tau_{k-j},\tau_k)\big), \tilde{Q}(0) - \eta + A^v(\tau_k) - \sum_{i=1}^{\eta} N_i^v(\tau_k)\Big).$$

As all jumps in $\tilde{Q}(t)$ on $[0,\gamma]$ occur at times $t \in \{\tau_k, k \geq 1\}$, and are of size 1,

$$\max_{j \in [0,k-1]} \big(A^v(\tau_{k-j},\tau_k) - \sum_{i=1}^{\eta} N_i^v(\tau_{k-j},\tau_k)\big) \leq 1 + \max_{j \in [0,k]} \big(A^v(\tau_{k-j},\tau_k) - \sum_{i=1}^{\eta} N_i^v(\tau_{k-j},\tau_k)\big).$$

It follows that

Corollary 7. *If $\tilde{Q}(0) = \eta$, then*

$$\tilde{Q}(\gamma) - \eta = \phi(v,\gamma,\eta).$$

In general,

$$\tilde{Q}(\gamma) - \eta \leq \max\Big(1 + \phi(v,\gamma,\eta), \phi(0,v,n) + A^v(\gamma) - \sum_{i=1}^{\eta} N_i^v(\gamma)\Big).$$

Before proceeding, it will be useful to prove a general comparison result for $G/G/n$ queues, which is very similar to Lemma 26 from Chapter 4. The key difference is that here we allow for both general initial conditions and differing numbers of servers. Recall that for an event $\{E\}$, $I(\{E\})$ denotes the indicator function of $\{E\}$. Then

Lemma 35. *Let Q^1 denote a FCFS $G/G/n^1$ queue, and Q^2 denote a FCFS $G/G/n^2$ queue, both with finite, strictly positive inter-arrival and processing times, s.t. $n^1 \geq n^2$. Let $Q^i(t)$ denote the number in system at time t in Q^i, and $\mathcal{L}^i \triangleq Q^i(0) - n^i$, $i \in \{1,2\}$. For $k \in \{1,\ldots,\mathcal{L}^i\}$, let T_k^i equal zero, and S_k^i denote the initial processing time of the kth job initially waiting in queue in Q^i, $i \in \{1,2\}$. For $k > \mathcal{L}^i$, let T_k^i denote the arrival time of the $(k - \mathcal{L}^i)$th arrival (after time 0) to Q^i, and S_k^i the processing time assigned to that job, $i \in \{1,2\}$. Further suppose that*

(i) $\mathcal{L}^1 = \mathcal{L}^2 \geq 0$, and we denote this common value by $\mathcal{L}$. Also, $S_k^1 = S_k^2$ for $k \in \{1,\ldots,\mathcal{L}\}$. That is, the kth job initially waiting in queue in Q^2 is assigned

the same processing time as the kth job initially waiting in queue in Q^1, $k = 1, \dots, \mathcal{L}$. In addition, we let W_i denote the initial processing time of the job initially being processed on server i in Q^1, $i \in \{1, \dots, n^1\}$.

(ii) *For each job J initially being processed in Q^1 on a server whose index belongs to the set $\{1, \dots, n^2\}$, there is a distinct corresponding job J' initially being processed in Q^2, s.t. the initial processing time of J in Q^1 equals the initial processing time of J' in Q^2.*

(iii) *$\{T_k^1, k \geq 1\}$ is a subsequence of $\{T_k^2, k \geq 1\}$.*

(iv) *For all $k > \mathcal{L}$, the job that arrives to Q^2 at time T_k^1 is assigned processing time S_k^1, the same processing time assigned to the job which arrives to Q^1 at that time.*

Then for all $t \geq 0$,

$$Q^1(t) \leq Q^2(t) + \sum_{i=n^2+1}^{n^1} I(W_i > t).$$

Proof. Let $Z_a^1(t)$ denote the number of jobs initially being processed in Q^1, on servers with index $i \leq n^2$, which are still in Q^1 at time t. Let $Z_b^1(t)$ denote the number of jobs initially being processed in Q^1, on servers with index $i > n^2$, which are still in Q^1 at time t. Let $Z^2(t)$ denote the number of jobs initially being processed in Q^2, which are still in Q^2 at time t. Note that by (ii), $Z_a^1(t) = Z^2(t)$ for all $t \geq 0$, and we denote this common value by $Z(t)$. Also, for all $t \geq 0$,

$$Z_b^1(t) = \sum_{i=n^2+1}^{n^1} I(W_i > t). \tag{5.2}$$

For $k \in \{1, \dots, \mathcal{L}\}$, let D_k^i denote the time at which the kth job initially waiting in queue in Q^i departs from Q^i, $i \in \{1, 2\}$. For $k > \mathcal{L}$, let D_k^i denote the time at which the job that arrives to $\mathcal{Q}^i$ at time T_k^1 departs from $\mathcal{Q}^i$, $i \in \{1, 2\}$. We now prove by induction that for $k \geq 1$, $D_k^2 \geq D_k^1$, from which the lemma follows. Observe that for all $k \geq 1$,

$$\begin{aligned} D_k^1 &= \inf\{t: \ t \geq T_k^1, Z(t) + Z_b^1(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n^1 - 1\} + S_k^1 \\ &= \inf\{t: \ t \geq T_k^1, Z(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n^1 - Z_b^1(t) - 1\} + S_k^1 \\ &\leq \inf\{t: \ t \geq T_k^1, Z(t) + \sum_{j=1}^{k-1} I(D_j^1 > t) \leq n^2 - 1\} + S_k^1 \qquad (5.3) \\ & \quad \text{since } Z_b^1(t) \leq n^1 - n^2 \text{ for all } t \geq 0, \text{ by (5.2).} \end{aligned}$$

Also,

$$D_k^2 \geq \inf\{t: \ t \geq T_k^1, Z(t) + \sum_{j=1}^{k-1} I(D_j^2 > t) \leq n^2 - 1\} + S_k^1, \qquad (5.4)$$

where the inequality in (5.4) arises since the job that arrives to $\mathcal{Q}^2$ at time T_k^1 may have to wait for additional jobs, which arrive at a time belonging to $\{T_k^2, k \geq 1\} \setminus \{T_k^1, k \geq 1\}$.

For the base case $k = 1$, note that $D_1^1 \leq \inf\{t : t \geq T_1^1, Z(t) \leq n^2 - 1\} + S_1^1$ by (5.3), while $D_1^2 \geq \inf\{t : t \geq T_1^1, Z(t) \leq n^2 - 1\} + S_1^1$ by (5.4).

Now assume the induction is true for all $j \leq k$. Then for all $t \geq 0$, $\sum_{j=1}^k I(D_j^2 > t) \geq \sum_{j=1}^k I(D_j^1 > t)$. Thus

$$\inf\{t: \ t \geq T_{k+1}^1, Z(t) + \sum_{j=1}^{k} I(D_j^1 > t) \leq n^2 - 1\} + S_{k+1}^1$$

$$\leq \inf\{t: \ t \geq T_{k+1}^1, Z(t) + \sum_{j=1}^{k} I(D_j^2 > t) \leq n^2 - 1\} + S_{k+1}^1.$$

It then follows from (5.3) and (5.4) that $D^1_{k+1} \leq D^2_{k+1}$, completing the induction. □

We now complete the proof of Theorem 27.

Proof of Theorem 27. Let us fix some $\gamma, \tau \in \mathbb{R}^+$ s.t. $\gamma \leq \tau$, and $\eta \in Z^+$ s.t. $\eta \leq n$. We begin by constructing Q and $\tilde{Q}^0_{\gamma,n}$ on the same probability space. Note that Q and $\tilde{Q}^0_{\gamma,n}$ have the same initial conditions. Namely, for $i = 1, \ldots, n$, there is a single job initially being processed on server i with initial processing time V^1_i, there are 0 jobs waiting in queue, and the time until the first arrival is U^1. We let $A(t)$ be the arrival process to Q. Let $\{t_k, k \geq 1\}$ denote the ordered sequence of event times in $A(t)$. It follows from the construction of $\tilde{A}^0_{\gamma,n}(t)$ that $\{t_k, k \geq 1\}$ is a subsequence of the set of event times in $\tilde{A}^0_{\gamma,n}(t)$. We let the processing time assigned to the arrival to $\tilde{Q}^0_{\gamma,n}$ at time t_k equal the processing time assigned to the arrival to Q at time t_k, $k \geq 1$. It follows that Q and $\tilde{Q}^0_{\gamma,n}$ satisfy the conditions of Lemma 35, and it follows from Corollary 6 that for all $x \geq 0$,

$$\mathbb{P}\big(Q(\tau) > x\big) \leq \mathbb{P}\big(\tilde{Q}^0_{\gamma,n}(\tau) > x\big). \tag{5.5}$$

It follows from Lemma 34 that at time γ, server i of $\tilde{Q}^0_{\gamma,n}$ is processing a job with remaining processing time (at time γ) $V^1_i(\gamma)$, $i = 1, \ldots, n$. It follows from Corollary 7 that at time γ, there are $\phi(0, \gamma, n)$ jobs waiting in queue in $\tilde{Q}^0_{\gamma,n}$. Also, the remaining time (at time γ) until the next arrival to $\tilde{Q}^0_{\gamma,n}$ is $U^1(\gamma)$. Thus by construction, the state of $\tilde{Q}^0_{\gamma,n}$ at time γ (if viewed as a Markov chain (see [4])) is identical to the state of Q^γ_n at time 0. It then follows from our construction, the Markov chain interpretation of the $GI/GI/n$ queue (see [4]), and Corollary 6 that we may construct $\tilde{Q}^0_{\gamma,n}$ and Q^γ_n on the same probability space s.t. $\tilde{Q}^0_{\gamma,n}(\gamma + s) = Q^\gamma_n(s)$ for all $s \geq 0$. Thus for all $x \geq 0$,

$$\mathbb{P}\big(\tilde{Q}^0_{\gamma,n}(\tau) > x\big) = \mathbb{P}\big(Q^\gamma_n(\tau - \gamma) > x\big). \tag{5.6}$$

We now construct Q^γ_n and Q^γ_η on the same probability space Ω. Suppose w.l.o.g. that $A(t)$, $\{N_i(t), i = 1, \ldots, n\}$, and all associated auxiliary r.v.s (e.g. $A^\gamma(t), \{N^\gamma_i(t), i = 1, \ldots, n\}, \{V^j_i(\gamma), i = 1, \ldots, n, j \geq 1\}$, $\phi(0, \gamma, n)$) have been constructed on Ω. Note that for $i \leq \eta$, the initial processing time of the job initially being processed on server i is $V^1_i(\gamma)$ in both Q^γ_n and Q^γ_η. Also, in both systems there are $\phi(0, \gamma, n)$ jobs initially waiting in queue, and we assign corresponding initial jobs waiting in queue the same processing time. We let $A^\gamma(t)$ be the arrival process to both systems, and assign the same processing time to corresponding arrivals. It follows that on Ω, Q^γ_n and Q^γ_η satisfy the conditions of Lemma 35, and

$$Q^\gamma_n(\tau - \gamma) \leq Q^\gamma_\eta(\tau - \gamma) + \sum_{i=\eta+1}^{n} I(V^1_i(\gamma) > \tau - \gamma). \tag{5.7}$$

Note that on Ω, all inter-arrival times for jobs arriving to Q^γ_η (except the first), and all processing times assigned to jobs not initially being processed in Q^γ_η, are independent of $\{V^1_i(\gamma), i = 1, \ldots, n\}$, and thus independent of $\sum_{i=\eta+1}^{n} I(V^1_i(\gamma) > \tau - \gamma)$. Also note that although our construction ensures that the arrival processes and assignment of processing times in both Q^γ_n and Q^γ_η coincide, we have not yet specified any particular construction for these arrival processes and processing times on Ω.

We now construct $\tilde{Q}^\gamma_{\tau-\gamma,\eta}$ on the same probability space Ω, and simultaneously give an explicit construction for the arrival process and assignment of processing times to jobs for Q^γ_η on Ω. Note that the initial conditions, arrival process, and assignment of processing times to jobs for $\tilde{Q}^\gamma_{\tau-\gamma,\eta}$ are all deterministic functions of $A(t)$, $\{N_i(t), i = 1, \ldots, n\}$, and the associated auxiliary r.v.s. It follows that $\tilde{Q}^\gamma_{\tau-\gamma,\eta}$, and the associated process $\tilde{Q}^\gamma_{\tau-\gamma,\eta}(t)$, have already been implicitly constructed on Ω. Note that by construction, Q^γ_η and $\tilde{Q}^\gamma_{\tau-\gamma,\eta}$ have the same initial conditions, and thus the same number $\phi(0, \gamma, n)$ of jobs initially waiting in queue. We assign corresponding initial jobs waiting in queue the same processing time, as dictated by the assignment of processing times to jobs in $\tilde{Q}^{\gamma}_{\tau-\gamma,\eta}$. Namely, if the kth job initially waiting in queue in $\tilde{Q}^{\gamma}_{\tau-\gamma,\eta}$ is the jth job assigned to server i in $\tilde{Q}^{\gamma}_{\tau-\gamma,\eta}$, it receives processing time $V_i^{j+1}(\gamma)$. We let $A^{\gamma}(t)$ be the arrival process to Q^{γ}_{η}. It follows from the construction of $\tilde{A}^{\gamma}_{\tau-\gamma,\eta}(t)$ that the sequence of arrival times to Q^{γ}_{η} is a subsequence of the set of arrival times to $\tilde{Q}^{\gamma}_{\tau-\gamma,\eta}$, and we assign the same processing time to corresponding arrivals. Namely, if the corresponding arrival in $\tilde{Q}^{\gamma}_{\tau-\gamma,\eta}$ is the jth job assigned to server i in $\tilde{Q}^{\gamma}_{\tau-\gamma,\eta}$, it receives processing time $V_i^{j+1}(\gamma)$. It follows that on the probability space Ω, Q^{γ}_{η} and $\tilde{Q}^{\gamma}_{\tau-\gamma,\eta}$ satisfy the conditions of Lemma 35, and it thus follows from (5.7) and Corollary 6 that

$$Q^{\gamma}_{n}(\tau-\gamma) \leq \tilde{Q}^{\gamma}_{\tau-\gamma,\eta}(\tau-\gamma) + \sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > \tau-\gamma). \tag{5.8}$$

Combining (5.5) - (5.8) with Corollary 7, and observing that γ, τ, η were general, we find that for all $t, x \geq 0$, $\mathbb{P}\big(Q(t) > x\big)$ is at most

$$\inf_{\substack{\gamma\in[0,t]\\ \eta\in[0,n]}} \mathbb{P}\Bigg(\max\Bigg(1+\phi(\gamma,t-\gamma,\eta), \phi(0,\gamma,n) + A^{\gamma}(t-\gamma) - \sum_{i=1}^{\eta} N_i^{\gamma}(t-\gamma)\Bigg) \tag{5.9}$$

$$+ \sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t-\gamma) \quad > \quad x-\eta\Bigg).$$

From definitions,

$$\begin{aligned}
\phi(\gamma, t-\gamma, \eta) &= \sup_{0\le s\le t-\gamma} \big(A(t-s,t) - \sum_{i=1}^{\eta} N_i(t-s,t)\big);\\
\phi(0,\gamma,n) &= \sup_{0\le s\le \gamma} \big(A(\gamma-s,\gamma) - \sum_{i=1}^{n} N_i(\gamma-s,\gamma)\big);\\
A^{\gamma}(t-\gamma) - \sum_{i=1}^{\eta} N_i^{\gamma}(t-\gamma) &= A(\gamma,t) - \sum_{i=1}^{\eta} N_i(\gamma,t);\\
\sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t-\gamma) &= \sum_{i=\eta+1}^{n} I(N_i(\gamma,t)=0).
\end{aligned}$$

It follows from elementary renewal theory (see [27]) that the joint distribution of

$$\phi(\gamma, t-\gamma, \eta), \phi(0,\gamma,n), A^{\gamma}(t-\gamma) - \sum_{i=1}^{\eta} N_i^{\gamma}(t-\gamma), \sum_{i=\eta+1}^{n} I(V_i^1(\gamma) > t-\gamma)$$

is the same as that of

$$\sup_{0\le s\le t-\gamma} \big(A(s) - \sum_{i=1}^{\eta} N_i(s)\big), \sup_{0\le s\le\gamma} \big(A^{t-\gamma}(s) - \sum_{i=1}^{n} N_i^{t-\gamma}(s)\big),$$

$$A(t-\gamma) - \sum_{i=1}^{\eta} N_i(t-\gamma), \sum_{i=\eta+1}^{n} I(N_i(t-\gamma)=0).$$

Combining the above with (5.9), and letting $\delta \triangleq t-\gamma$, completes the proof of the first part of the theorem.

We now prove the corresponding steady-state result. Note that for all $\delta \le t$,

$$\sup_{\delta\le s\le t} \big(A(s) - \sum_{i=1}^{n} N_i(s)\big) \le \sup_{s\ge\delta} \big(A(s) - \sum_{i=1}^{n} N_i(s)\big).$$

It follows that for any fixed values of η and x, $\mathbb{P}\big(Q(t) > x\big)$ is at most

$$\inf_{\delta\in[0,t]} \mathbb{P}\Bigg(\max\Bigg(1 + \sup_{0\leq s\leq\delta}\big(A(s) - \sum_{i=1}^{\eta} N_i(s)\big), \sup_{s\geq\delta}\big(A(s) - \sum_{i=1}^{n} N_i(s)\big) + \sum_{i=\eta+1}^{n} N_i(\delta)\Bigg) \tag{5.10}$$
$$+ \sum_{i=\eta+1}^{n} I(N_i(\delta) = 0) \ > \ x - \eta\Bigg).$$

Since (5.10) is monotone decreasing in t, the corresponding limit exists as $t \rightarrow \infty$, completing the proof. $\square$

5.4 Asymptotic Bound in the Halfin-Whitt Regime

In this section, we use Theorem 27 to bound the FCFS $GI/GI/n$ queue in the H-W regime, by proving an asymptotic analogue of Theorem 27. Suppose that the H-W and T_0 assumptions hold. Recall that a Gaussian process on $\mathbb{R}$ is a stochastic process $Z(t)_{t\geq 0}$ s.t. for any finite set of times $t_1, \ldots, t_k$, the vector $\big(Z(t_1), \ldots, Z(t_k)\big)$ has a Gaussian distribution. A Gaussian process $Z(t)$ is known to be uniquely determined by its mean function $\mathbb{E}[Z(t)]$ and covariance function $\mathbb{E}[Z(s)Z(t)]$, and refer the reader to [35],[55],[2],[77], and the references therein for details on existence, continuity, etc. It is proven in [110] Theorem 2 that there exists a continuous Gaussian process $\mathcal{D}(t)$ s.t. $\mathbb{E}[\mathcal{D}(t)] = 0, \mathbb{E}[\mathcal{D}(s)\mathcal{D}(t)] = \mathbb{E}[\big(N_1(s) - \mu s\big)\big(N_1(t) - \mu t\big)]$ for all $s, t \geq 0$. Let $\mathcal{A}(t)$ denote the continuous Gaussian process s.t. $\mathbb{E}[\mathcal{A}(t)] = 0, \mathbb{E}[\mathcal{A}(s)\mathcal{A}(t)] = \mu c_A^2 \min(s,t)$, namely $\mathcal{A}(t)$ is a driftless Brownian motion. Let $\mathcal{Z}(t) \triangleq \mathcal{A}(t) - \mathcal{D}(t)$, where $\mathcal{A}(t)$ and $\mathcal{D}(t)$ are independent. Then our main asymptotic upper bound is that

Theorem 28. *For all* $B > 0$ *and* $x \in \mathbb{R}$, $\limsup_{n\rightarrow\infty} \mathbb{P}\Big(\big(Q^n_B(\infty) - n\big)n^{-\frac{1}{2}} > x\Big)$ *is* *at most*

$$\inf_{\substack{\delta\geq 0\\ \eta\geq 0}} \mathbb{P}\Bigg(\max\bigg(\sup_{0\leq t\leq\delta} \big(\mathcal{Z}(t)+(\eta-B)\mu t\big)\ \ ,\ \ \sup_{t\geq\delta}\big(\mathcal{Z}(t)-B\mu t\big)\ +\ \eta\mu\delta\bigg) \geq x+\eta\mathbb{P}\big(R(X)\leq\delta\big)\Bigg).$$

Let $A^n_B(t)$ denote an equilibrium renewal process with renewal distribution $A\lambda^{-1}_{n,B}$, independent of $\{N_i(t), i=1,\ldots,n\}$, $Z^n_B(t) \triangleq A^n_B(t)-\sum_{i=1}^n N_i(t)$, and $Z^n_B(s,t) \triangleq Z^n_B(t)-Z^n_B(s)$. Also, we define $f_n(x) \triangleq \lfloor n - xn^{\frac{1}{2}}\rfloor$. Then it follows from Theorem 27 that for all $x\in\mathbb{R}$, $\mathbb{P}\Big(\big(Q^n_B(\infty)-n\big)n^{-\frac{1}{2}}>x\Big)$ is at most the infimum, taken over all $\delta\geq 0$ and $\eta\in[0,n^{\frac{1}{2}}]$, of the probability of the event

$$\Bigg\{\max\bigg(1+\sup_{0\leq t\leq\delta}\big(Z^n_B(t)+\sum_{i=f_n(\eta)+1}^{n} N_i(t)\big), \sup_{t\geq\delta} Z^n_B(t)+\sum_{i=f_n(\eta)+1}^{n} N_i(\delta)\bigg) + \sum_{i=f_n(\eta)+1}^{n} I(N_i(\delta)=0)\ >\ f_n(-x)-f_n(\eta)\Bigg\}. \tag{5.11}$$

Before completing the proof of Theorem 28, we establish some preliminary weak convergence results to aid in the analysis of (5.11).

5.4.1 Preliminary weak convergence results

For an excellent review of weak convergence, and the associated spaces (e.g. $D[0,T]$) and topologies/metrics (e.g. uniform, J_1), the reader is referred to [113]. We now review several results from Chapter 4. It is proven in Chapter 4, Lemma 16 that

Theorem 29. *For any $B>0$ and $T\in[0,\infty)$, the sequence of processes $\{n^{-\frac{1}{2}}Z^n_B(t)_{0\leq t\leq T}, n\geq 1\}$ converges weakly to $\big(\mathcal{Z}(t)-B\mu t\big)_{0\leq t\leq T}$ in the space $D[0,T]$ under the J_1 topology.*

The behavior of $n^{-\frac{1}{2}}Z_B^n(t)$ over unbounded time intervals is also addressed in Chapter 4. In particular, Chapter 4, Equations 4.25 - 4.26 establish that

Theorem 30. *For all* $B > 0$, $\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}Z_B^n(T) > -\frac{B}{2}\mu T\Big) = 0$; *and* $\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq 0}Z_B^n(t) > T\Big) = 0.$

We now combine Theorems 29 - 30 to establish that

Corollary 8. *For any fixed* $B > 0$, $\delta \geq 0$, *and* $\epsilon > 0$, *there exists* $T \in (\delta, \infty)$, *depending only on* B, δ, *and* ϵ, *s.t.*

$$\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{t\geq T}Z_B^n(t) \geq \sup_{t\geq\delta}Z_B^n(t)\Big) < \epsilon.$$

Proof. It follows from the monotonicity of the supremum operator and a union bound that the l.h.s. of (8) is at most

$$\limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}Z_B^n(T) > -\frac{B}{2}\mu T\Big) + \limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}\sup_{t\geq 0}Z_B^n(T, T+t) > \frac{B}{8}\mu T\Big)$$
$$+\limsup_{n\to\infty}\mathbb{P}\Big(n^{-\frac{1}{2}}Z_B^n(\delta) \leq -\frac{B}{8}\mu T\Big).$$

The corollary then follows from the stationary increments property of $Z_B^n(t)$, Theorem 29, and Theorem 30. □

It is proven in [110] Theorem 2 that

Theorem 31. *For any* $T \in [0, \infty)$, *the sequence of processes* $\{n^{-\frac{1}{2}}(\sum_{i=1}^n N_i(t) - n\mu t)_{0\leq t\leq T}, n \geq 1\}$ *converges weakly to* $\mathcal{D}(t)_{0\leq t\leq T}$ *in the space* $D[0, T]$ *under the* J_1 *topology.*

Note that

$$\lim_{n\to\infty} n^{-\frac{1}{2}}\big(n - f_n(\eta)\big) = \eta, \quad \text{and} \quad \lim_{n\to\infty} n^{-\frac{1}{2}}\big(f_n(-x) - f_n(\eta)\big) = x + \eta. \tag{5.12}$$

It follows from Theorem 31 and (5.12) that

Corollary 9. *For all $\eta \geq 0$ and $T \in [0,\infty)$, the sequence of processes $\left\{n^{-\frac{1}{2}}\left(\sum_{i=f_n(\eta)+1}^{n} N_i(t) - \big(n - f_n(\eta)\big)\mu t\right)_{0\leq t\leq T}, n \geq 1\right\}$ converges weakly to the constant process 0 in the space $D[0,T]$ under the J_1 topology.*

Finally, we establish the convergence of several additional sequences.

Lemma 36. *The sequence of r.v.s $\left\{n^{-\frac{1}{2}}\sum_{i=f_n(\eta)+1}^{n} N_i(\delta), n \geq 1\right\}$ converges weakly to the constant $\eta\mu\delta$, and the sequence of r.v.s $\left\{n^{-\frac{1}{2}}\sum_{i=f_n(\eta)+1}^{n} I(N_i(\delta) = 0), n \geq 1\right\}$ converges weakly to the constant $\eta\mathbb{P}\big(R(X) > \delta\big)$.*

Proof. Both convergences follow from the strong law of large numbers. □

5.4.2 Proof of Theorem 28

We now complete the proof of Theorem 28

Proof of Theorem 28. Note that it suffices to demonstrate that for each fixed $B > 0$ and $\delta, \eta \geq 0$, the $\limsup_{n\to\infty}$ of the probability of event (5.11) is at most

$$\mathbb{P}\left(\max\left(\sup_{0\leq t\leq\delta}\big(\mathcal{Z}(t)+(\eta-B)\mu t\big)\ ,\ \sup_{t\geq\delta}\big(\mathcal{Z}(t)-B\mu t\big)+\eta\mu\delta\right)\ \geq\ x+\eta\mathbb{P}\big(R(X)\leq\delta\big)\right). \tag{5.13}$$

Let us fix some $B > 0$ and $\delta, \eta \geq 0$. Applying Corollary 8 and a union bound, and multiplying (5.11) through by $n^{-\frac{1}{2}}$, it thus suffices to demonstrate that

$$\lim_{T\to\infty}\limsup_{n\to\infty}\mathbb{P}\Bigg(n^{-\frac{1}{2}}\max\Bigg(1+\sup_{0\leq t\leq\delta}\Big(Z^n_B(t)+\sum_{i=f_n(\eta)+1}^{n}N_i(t)\Big), \tag{5.14}$$
$$\sup_{t\in[\delta,T]}Z^n_B(t)+\sum_{i=f_n(\eta)+1}^{n}N_i(\delta)\Bigg)$$
$$+n^{-\frac{1}{2}}\sum_{i=f_n(\eta)+1}^{n}I(N_i(\delta)=0)-n^{-\frac{1}{2}}\big(f_n(-x)-f_n(\eta)\big)\geq 0\Bigg)$$

is at most (5.13). For any fixed T, it follows from Theorem 29, Corollary 9, Lemma 36, and the continuity of the supremum map in the space $D[0,T]$ under the J_1 topology (see [113] Theorem 13.4.1) that

$$\left\{n^{-\frac{1}{2}} \max\left(1 + \sup_{0 \le t \le \delta} \left(Z^n_B(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(t)\right), \sup_{\delta \le t \le T} Z^n_B(t) + \sum_{i=f_n(\eta)+1}^{n} N_i(\delta)\right), n \ge 1\right\}$$

converges weakly to $\max\left(\sup_{0\le t\le\delta} \left(\mathcal{Z}(t) + (\eta - B)\mu t\right)\ ,\ \sup_{\delta \le t \le T} \left(\mathcal{Z}(t) - B\mu t\right) + \eta\mu\delta\right)$. That (5.14) is at most (5.13) then follows from Lemma 36, (5.12), and the Portmanteau Theorem (see [7]), which states that a sequence of r.v.s $\{X_n\}$ converges weakly to the r.v. X_∞ iff for all closed subsets C of $\mathbb{R}$, $\limsup_{n\to\infty} \mathbb{P}(X_n \in C) \le \mathbb{P}(X_\infty \in C)$ iff for all open subsets O of $\mathbb{R}$, $\mathbb{P}(X_\infty \in O) \le \liminf_{n\to\infty} \mathbb{P}(X_n \in O)$. □

5.5 Proof of Bound for Probability of Delay as $B \to \infty$

In this section we complete the proof of Theorem 24. We begin by recalling some additional results from Chapter 4.

Lemma 37. $\lim_{t\to\infty} \mathbb{E}[\left(t^{-\frac{1}{2}}\mathcal{Z}(t)\right)^2] = \mu(c_A^2 + c_S^2)$. *Also, the r.v.* $\sup_{t\ge0}\left(\mathcal{Z}(t) - \frac{B}{2}\mu t\right)$ *is a.s. finite.*

Proof. The first part of the lemma follows from the proof of Chapter 4, Corollary 5. The second part of the lemma follows from the proof of Chapter 4, Theorem 12, specifically the discussion after Equation (30). □

We now prove a modified variant of Theorem 28, which has the interpretation of setting $\delta = \infty, \eta = \frac{B}{2}$ in Theorem 28.

Corollary 10. *For all $B > 0$ and $x \in \mathbb{R}$, $\limsup_{n\to\infty} \mathbb{P}\Big(\big(Q^n_B(\infty) - n\big)n^{-\frac{1}{2}} > x\Big)$ is at most*

$$\mathbb{P}\Big(\sup_{t\geq 0} \big(\mathcal{Z}(t) - \frac{B}{2}\mu t\big) \ \geq \ x + \frac{B}{2}\mu\Big).$$

Proof. It follows from setting $\eta = \frac{B}{2}\mu$ in Theorem 28, combined with the monotonicity of the supremum operator and a union bound, that for all $B > 0$ and $x \in \mathbb{R}$, $\limsup_{n\to\infty} \mathbb{P}\Big(\big(Q^n_B(\infty) - n\big)n^{-\frac{1}{2}} > x\Big)$ is at most

$$\inf_{\delta\geq 0}\Bigg(\mathbb{P}\Big(\sup_{t\geq 0} \big(\mathcal{Z}(t) - \frac{B}{2}\mu t\big) \ \geq \ x + \frac{B}{2}\mu\mathbb{P}\big(R(X) \leq \delta\big)\Big) \tag{5.15}$$

$$+\mathbb{P}\Big(\sup_{t\geq \delta} \big(\mathcal{Z}(t) - B\mu t\big) \geq -\frac{B}{2}\mu\delta\Big)\Bigg). \tag{5.16}$$

Let $\mathcal{Z}(s,t) = \mathcal{Z}(t) - \mathcal{Z}(s)$. Then it follows from the fact that $\mathcal{Z}(t)$ has stationary increments and a union bound that (5.16) is at most

$$\mathbb{P}\Big(\mathcal{Z}(\delta) \geq \frac{B}{4}\mu\delta\Big) + \mathbb{P}\Big(\sup_{t\geq 0} \big(\mathcal{Z}(t) - B\mu t\big) \geq \frac{B}{4}\mu\delta\Big). \tag{5.17}$$

It follows from Lemma 37 that for any $\epsilon > 0$, there exists $\delta_\epsilon < \infty$ s.t. for all $\delta \geq \delta_\epsilon$, (5.17) is at most ϵ. The corollary then follows from (5.15) and the continuity of probability measures, since $\mathbb{P}\big(R(X) \leq \delta\big) \to 1$ as $\delta \to \infty$. □

Finally, we state a well-known result from the theory of Gaussian processes (see [2] Inequality 2.4) which will be critical to our proof.

Lemma 38. *Let $\mathcal{X}(t)$ denote any centered, continuous Gaussian process, and T any bounded interval of $\mathbb{R}^+$. Let $\sigma^2_T \overset{\Delta}{=} \sup_{t\in T}\mathbb{E}[\mathcal{X}^2(t)]$, and suppose $\sigma^2_T < \infty$. Then for* *all $\epsilon > 0$, there exists M_ϵ, depending only on $\mathcal{X}(t), T$, and ϵ, s.t. for all $x > M_\epsilon$,*

$$\mathbb{P}\big(\sup_{t\in T}\mathcal{X}(t) > x\big) \leq \exp\Bigg(- \bigg((2\sigma_T^2)^{-1} - \epsilon\bigg)x^2\Bigg).$$

We now complete the proof of Theorem 24

Proof of Theorem 24. It follows from Corollary 10 and a union bound that for all $B > 0$, $\limsup_{n\to\infty}\mathbb{P}\bigg(Q_B^n(\infty) \geq n\bigg)$ is at most

$$\sum_{k=0}^{\infty}\mathbb{P}\bigg(\sup_{k\leq t\leq k+1}\big(\mathcal{Z}(t) - \frac{B}{2}\mu t\big) \quad\geq\quad \frac{B}{2}\mu\bigg). \tag{5.18}$$

Note that for any fixed $k \geq 0$, $\mathbb{P}\bigg(\sup_{k\leq t\leq k+1}\big(\mathcal{Z}(t) - \frac{B}{2}\mu t\big) \quad\geq\quad \frac{B}{2}\mu\bigg)$ equals

$$\mathbb{P}\Bigg(\big(\mathcal{Z}(k) - \frac{B}{4}\mu k\big) + \sup_{k\leq t\leq k+1}\bigg(\big(\mathcal{Z}(t) - \frac{B}{2}\mu t\big) - \big(\mathcal{Z}(k) - \frac{B}{2}\mu k\big) - \frac{B}{4}\mu k\bigg) \geq \frac{B}{2}\mu\Bigg).$$

It then follows from the stationary increments property of $\mathcal{Z}(t)$ and a union bound that (5.18) is at most

$$\sum_{k=0}^{\infty}\mathbb{P}\bigg(\mathcal{Z}(k) \geq \frac{B}{4}\mu(k+1)\bigg) \tag{5.19}$$

$$+\sum_{k=0}^{\infty}\mathbb{P}\bigg(\sup_{0\leq t\leq 1}\mathcal{Z}(t) \geq \frac{B}{4}\mu(k+1)\bigg). \tag{5.20}$$

Since $\mathcal{Z}(0) = 0$, and $\frac{B}{4}\mu(k+1) \geq \frac{B}{4}\mu k$, (5.19) is at most

$$\sum_{k=1}^{\infty}\mathbb{P}\bigg(k^{-\frac{1}{2}}\mathcal{Z}(k) \geq \frac{B}{4}\mu k^{\frac{1}{2}}\bigg). \tag{5.21}$$

It follows from Lemma 37 that $c \triangleq \sup_{k\geq 1}\mathbb{E}\big[\Big(k^{-\frac{1}{2}}\mathcal{Z}(k)\Big)^2\big] < \infty$ is a finite, strictly positive constant depending only on A and S. Let G denote a normally distributed r.v. with mean 0 and variance 1. Note that for all $x \geq 1$,

$$\begin{aligned}\mathbb{P}(G > x) &= (2\pi)^{-\frac{1}{2}}\int_x^{\infty}\exp(-\frac{y^2}{2})dy \\ &\leq (2\pi)^{-\frac{1}{2}}\int_x^{\infty} y\exp(-\frac{y^2}{2})dy \;=\; (2\pi)^{-\frac{1}{2}}\exp(-\frac{x^2}{2}).\end{aligned}$$

Combining the above, we find that for all $B \geq 4c^{\frac{1}{2}}\mu^{-1}$, (5.19) is at most

$$\sum_{k=1}^{\infty}\exp\Big(-\frac{\mu^2}{32c}B^2k\Big) = \frac{\exp\Big(-\frac{\mu^2}{32c}B^2\Big)}{1-\exp\Big(-\frac{\mu^2}{32c}B^2\Big)}. \tag{5.22}$$

We now bound (5.20). Let $c_2 \triangleq \sup_{t\in[0,1]}\mathbb{E}[\mathcal{Z}^2(t)]$, and note that trivially $c_2 < \infty$. It follows from Lemma 38 that there exists $B_0 < \infty$, depending only on A and S, s.t. $B \geq B_0$ implies that for all $k \geq 1$,

$$\mathbb{P}(\sup_{0\leq t\leq 1}\mathcal{Z}(t) > \frac{B}{4}\mu k) \leq \exp\big(-(4c_2)^{-1}(\frac{B}{4}\mu k)^2\big).$$

It follows that (5.20) is at most

$$\begin{aligned}\sum_{k=1}^{\infty}\exp\big(-\frac{\mu^2}{64c_2}B^2k^2\big) &\leq \sum_{k=1}^{\infty}\exp\big(-\frac{\mu^2}{64c_2}B^2k\big) \\ &= \frac{\exp\big(-\frac{\mu^2}{64c_2}B^2\big)}{1-\exp\big(-\frac{\mu^2}{64c_2}B^2\big)}.\end{aligned} \tag{5.23}$$

Using (5.22) to bound (5.19) and (5.23) to bound (5.20) completes the proof. □

5.6 Proof of Bound for Probability of Delay as $B \to 0$

In this section we complete the proof of Theorem 25. We proceed by carefully analyzing the covariance structure of $\mathcal{Z}(t)$, so that we may apply known results from the theory of Gaussian processes, in particular the well-known Slepian's lemma (see [49]) for comparing the suprema of multivariate Gaussian r.v.s., to prove our main results.

5.6.1 Slepian's lemma

We now formally state a particular variant of Slepian's lemma, proven in [49] Theorem 1.1.

Theorem 32. *For $k \geq 1$, let $\big(X_1, \ldots, X_k\big)$, $\big(Y_1, \ldots, Y_k\big)$ denote two zero-mean multivariate Gaussian r.v.s, each in $\mathbb{R}^k$. Further suppose that $\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2], i = 1, \ldots, k$, and $\mathbb{E}[X_i X_j] \geq \mathbb{E}[Y_i Y_j]$ for all $i, j \in \{1, \ldots, k\}$. Then for all vectors $(z_1, \ldots, z_k) \in \mathbb{R}^k$,*

$$\mathbb{P}\big(\bigcap_{i=1,\ldots,k} \{X_i \leq z_i\}\big) \geq \mathbb{P}\big(\bigcap_{i=1,\ldots,k} \{Y_i \leq z_i\}\big).$$

We now restate Theorem 32 for continuous Gaussian processes over an interval, which follows from Theorem 32, continuity, and the separability of $\mathbb{R}$ (see [82]).

Corollary 11. *Let $T \triangleq [t_1, t_2]$ denote any closed interval of $\mathbb{R}^+$. Let $\mathcal{X}(t)$ and $\mathcal{Y}(t)$ denote any two continuous zero-mean Gaussian processes s.t. $\mathbb{E}[\mathcal{X}^2(t)] = \mathbb{E}[\mathcal{Y}^2(t)]$ for all $t \in T$, and $\mathbb{E}[\mathcal{X}(s)\mathcal{X}(t)] \geq \mathbb{E}[\mathcal{Y}(s)\mathcal{Y}(t)]$ for all $s, t \in T$. Let $g(t)$ denote any function which is continuous on T. Then*

$$\mathbb{P}\bigg(\sup_{t \in T} \big(\mathcal{X}(t) - g(t)\big) \leq 0\bigg) \geq \mathbb{P}\bigg(\sup_{t \in T} \big(\mathcal{Y}(t) - g(t)\big) \leq 0\bigg).$$

We also state another variant of Theorem 32 for Gaussian processes, which will be useful in our analysis.

Corollary 12. *Let $T^1 \triangleq [t_1, t_2]$, $T^2 \triangleq [t_2, t_3]$ denote any closed intervals of $\mathbb{R}^+$ which intersect at exactly one point. Let $\mathcal{X}(t)$ denote any continuous zero-mean Gaussian processes s.t. $\mathbb{E}[\mathcal{X}(s)\mathcal{X}(t)] \geq 0$ for all $s, t \geq 0$. Let $g(t)$ denote any function which is continuous on $T^1 \bigcup T^2$. Then*

$$\mathbb{P}\Big(\sup_{t \in T^1 \bigcup T^2} (\mathcal{X}(t) - g(t)) \leq 0\Big) \geq \mathbb{P}\Big(\sup_{t \in T^1} (\mathcal{X}(t) - g(t)) \leq 0\Big)\mathbb{P}\Big(\sup_{t \in T^2} (\mathcal{X}(t) - g(t)) \leq 0\Big).$$

Proof. Consider some set $t_1, \ldots, t_k$ of times belonging to the interval T^1, and a set $t_{k+1}, \ldots, t_{2k}$ of times belonging to the interval T^2. Let $(Y_1, \ldots, Y_{2k}) \in \mathbb{R}^{2k}$ denote the zero-mean multivariate Gaussian r.v. s.t. $\mathbb{E}[Y_i^2] = \mathbb{E}[\mathcal{X}^2(t_i)], i = 1, \ldots, 2k$, $\mathbb{E}[Y_i Y_j] = \mathbb{E}[\mathcal{X}(t_i)\mathcal{X}(t_j)], 1 \leq i \leq j \leq k$, $\mathbb{E}[Y_i Y_j] = \mathbb{E}[\mathcal{X}(t_i)\mathcal{X}(t_j)], k+1 \leq i \leq j \leq 2k$, and $\mathbb{E}[Y_i Y_j] = 0$ otherwise (i.e. $i \in \{1, \ldots, k\}$, $j \in \{k+1, \ldots, 2k\}$). That such a multivariate Gaussian exists follows from the fact that $(Y_1, \ldots, Y_k)$ is distributed as $\big(\mathcal{X}(t_1), \ldots, \mathcal{X}(t_k)\big)$, $(Y_{k+1}, \ldots, Y_{2k})$ is distributed as $\big(\mathcal{X}(t_{k+1}), \ldots, \mathcal{X}(t_{2k})\big)$, and $(Y_1, \ldots, Y_k)$ is independent of $(Y_{k+1}, \ldots, Y_{2k})$. Then we may apply Theorem 32 to find that

$$\begin{aligned} \mathbb{P}(\bigcap_{i=1,\ldots,2k} \{\mathcal{X}(t_i) \leq g(t_i)\}) &\geq \mathbb{P}(\bigcap_{i=1,\ldots,2k} \{Y_i \leq g(t_i)\}) \\ &= \mathbb{P}(\bigcap_{i=1,\ldots,k} \{Y_i \leq g(t_i)\})\mathbb{P}(\bigcap_{i=k+1,\ldots,2k} \{Y_i \leq g(t_i)\}). \end{aligned}$$

The corollary then follows by continuity and separability, by letting $\{t_i, i = 1, \ldots, 2k\}$ become dense in $T^1 \bigcup T^2$. □

5.6.2 Properties of Brownian motion, the Ornstein-Uhlenbeck process, and the three-dimensional Bessel process.

In this subsection we review several properties of Brownian motion, the Ornstein-Uhlenbeck process, and the three-dimensional Bessel process. For a r.v. X, let $V[X]$ denote the variance of X. For r.v.s X, Y, let $V[X,Y] \overset{\Delta}{=} \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ denote the covariance of X and Y.

Brownian motion

For $b > 0$, let $\mathcal{B}^b(t)$ denote a Brownian motion initialized to b; namely, the continuous Gaussian process s.t. $\mathbb{E}[\mathcal{B}^b(t)] = b$, $V[\mathcal{B}^b(s), \mathcal{B}^b(t)] = s$ for all $0 \leq s \leq t$. We now state several basic properties of Brownian motion, and refer the reader to [8] for details. Recall that G is a normally distributed r.v. with mean 0 and variance 1. For a stochastic process $Z(t)$, let τ_Z^a denote the first hitting time of $Z(t)$ to a, where $\tau_Z^a = \infty$ if no such time exists.

Theorem 33. *Brownian motion has the following properties.*

(i) For all $t \geq 0$, $\mathbb{P}\big(\sup_{0 \leq s \leq t} \mathcal{B}^0(s) > x\big) = 2\mathbb{P}\big(G > xt^{-\frac{1}{2}}\big)$.

(ii) For all $c_1, c_2 > 0$, $\mathbb{P}\big(\tau_{\mathcal{B}^0}^{c_1} < \tau_{\mathcal{B}^0}^{-c_2}\big) = \frac{c_2}{c_1 + c_2}$.

(iii) For all $c, x > 0$, $\mathbb{P}\Bigg(\sup_{t \geq 0}\big(\mathcal{B}^0(t) - ct\big) > x\Bigg) = \exp(-2cx)$.

We also mention one other relevant result, namely an interesting independence that arises when studying functions of Brownian motion and its supremum. Let e_1 denote an exponentially distributed r.v. with mean 1, namely $\mathbb{P}(e_1 > x) = \exp(-x)$ for all $x > 0$. Let $\{\mathcal{B}_i^b(t), b \in \mathbb{R}, i \geq 1\}$ denote a collection of independent Brownian motions, with $\mathcal{B}_i^b(t)$ initialized to b. Then it follows from the main result of [96] (see also [117]) that

Theorem 34. *For all* $t \geq 0$, $\sup_{0 \leq s \leq t} \mathcal{B}_1^0(s)\big(\sup_{0 \leq s \leq t} \mathcal{B}_1^0(s) - \mathcal{B}_1^0(t)\big)$ *has the same distribution as* $\frac{t}{2}e_1$, *and is* independent *of* $\mathcal{B}_1^0(t)$.

Ornstein-Uhlenbeck process

For any $\rho > 0$, let $\mathcal{U}^\rho(t)$ denote the centered stationary Ornstein-Uhlenbeck (O-U) process whose correlations decay exponentially (over time) at rate ρ. Namely $\mathcal{U}^\rho(t)$ is the continuous Gaussian process s.t. $\mathbb{E}[\mathcal{U}^\rho(t)] = 0, V[\mathcal{U}^\rho(s), \mathcal{U}^\rho(t)] = \exp\big(-\rho(t-s)\big)$ for all $0 \leq s \leq t$. For a review of the basic properties of O-U processes (e.g. existence, continuity), we refer the reader to [34]. A law of the iterated logarithm is known to hold for the O-U process, and we now state a particular variant, which follows from a more general result for continuous, stationary, Gaussian processes proven in [78].

Theorem 35. *For any fixed* $\rho, \epsilon > 0$, *one may construct* $\mathcal{U}^\rho(t)$ *on the same probability space as an a.s. finite r.v.* Z, *whose distribution depends only on* ρ *and* ϵ, *s.t.* $|\mathcal{U}^\rho(t)| \leq Z + (1+\epsilon)\big(2\log(t+1)\big)^{\frac{1}{2}}$ *for all* $t \geq 0$.

Three-dimensional Bessel process

For any $b > 0$, let $\mathcal{S}^b(t)$ denote the three-dimensional Bessel process initialized to b. We now formally define $\mathcal{S}^b(t)$ as the solution to a certain stochastic integral equation. The stochastic integral equation

$$X_t = b^2 + 3t + 2\int_0^t |X_t|^{\frac{1}{2}} dB_s \tag{5.24}$$

has a unique strong solution $\mathcal{X}^{b^2}(t)$, which is non-negative; we refer the reader to the survey paper [48] for details. Then $\mathcal{S}^b(t)$, the three-dimensional Bessel process initialized to b, is defined as $\big(\mathcal{X}^{b^2}(t)\big)^{\frac{1}{2}}$. Then it is well-known that $\mathcal{S}^0(t)$, namely the three-dimensional Bessel process initialized to 0, is distributed as $\big(\sum_{i=1}^3 \mathcal{B}_i^0(t)^2\big)^{\frac{1}{2}}$, namely the radial distance process of a three-dimensional Brownian motion. An elegant construction for $\mathcal{S}^b(t)$ is given in [115], where it is shown that for $b > 0$, $\mathcal{S}^b(t)$ is distributed as the 'gluing together' of two Brownian motions initialized to b, and a three-dimensional Bessel process initialized to 0. Let $\{U_b, b \geq 0\}$ denote a set of independent uniformly distributed r.v.s, where U_b has the uniform distribution on $[0, b]$. Let $\{\mathcal{S}_1^b(t), b \geq 0\}$ denote a set of independent three-dimensional processes, where $\mathcal{S}_1^b(t)$ is initialized to b. Suppose $\{\mathcal{B}_i^b(t), b \in \mathbb{R}, i \geq 1\}$, $\{U_b, b \geq 0\}$, and $\{\mathcal{S}_1^b(t), b \geq 0\}$ are mutually independent. Then it is proven in [115] Theorem 3.1 that

Theorem 36. *For $b > 0$, define*

$$\mathcal{X}(t) \triangleq \begin{cases} \mathcal{B}_1^b(t) & 0 \leq t < \tau_{\mathcal{B}_1^b}^{U_b} \\ \mathcal{B}_2^b\left(\tau_{\mathcal{B}_1^b}^{U_b} + \tau_{\mathcal{B}_2^b}^{U_b} - t\right) & \tau_{\mathcal{B}_1^b}^{U_b} \leq t < \tau_{\mathcal{B}_1^b}^{U_b} + \tau_{\mathcal{B}_2^b}^{U_b} \\ \mathcal{S}_1^0\left(t - \tau_{\mathcal{B}_1^b}^{U_b} - \tau_{\mathcal{B}_2^b}^{U_b}\right) & \tau_{\mathcal{B}_1^b}^{U_b} + \tau_{\mathcal{B}_2^b}^{U_b} \leq t < \infty \end{cases}$$

Then the distribution of the process $\mathcal{X}(t)$ is identical to the distribution of the process $\mathcal{S}^b(t)$.

It is also proven in [115] that for $0 < b < c < \infty$, many additional distributional relationships hold between the process $\mathcal{B}^b(t)$ conditioned to hit c before 0, i.e. the event $\left\{\tau_{\mathcal{B}^b}^c < \tau_{\mathcal{B}^b}^0\right\}$, and the process $\mathcal{S}^b(t)$; we refer the reader to [115] for details. A particularly relevant result, proven in [115] and restated in [84] Proposotion 1.1, is that

Theorem 37. *For any fixed $0 < b < c < \infty$, the conditional distribution of the r.v.*

$$\tau_{\mathcal{B}^b}^c \quad \textit{given} \quad \left\{\tau_{\mathcal{B}^b}^c < \tau_{\mathcal{B}^b}^0\right\}$$

is identical to the distribution of the r.v.

$$\tau^c_{\mathcal{S}^b}.$$

Also, the conditional distribution of the process

$$\mathcal{B}^b(t)_{0 \leq t \leq \tau^c_{\mathcal{B}^b}} \quad \textit{given} \quad \{\tau^c_{\mathcal{B}^b} < \tau^0_{\mathcal{B}^b}\}$$

is identical to the distribution of the process

$$\mathcal{S}^b(t)_{0 \leq t \leq \tau^c_{\mathcal{S}^b}}.$$

A law of the iterated logarithm is known to hold for the three-dimensional Bessel process, and we now state a particular variant, which follows from [53] Theorem 2 and continuity.

Theorem 38. *For any $b, \epsilon > 0$, one may construct $\mathcal{S}^b(t)$ on the same probability space as an a.s. finite r.v. Z, whose distribution depends only on b and ϵ, s.t. $\mathcal{S}^b(t) \geq t^{\frac{1}{2}-\epsilon} - Z$ for all $t \geq 0$.*

The hitting times of $\mathcal{S}^b(t)$ are well-studied [67],[47],[83],[16],[15]. A relevant result, whose proof we include for completeness, is that

Theorem 39. *For all $b, \epsilon > 0$, there exists $T, \delta \in (0, \infty)$, depending only on b and ϵ, s.t. $M > T$ implies*

$$\mathbb{P}\big(\tau^M_{\mathcal{S}^b} < \delta M^2\big) \leq \epsilon.$$

Proof. The proof is deferred to the appendix. □

5.6.3 Renewal theory

We now review some results from renewal theory, which will be necessary in analyzing the covariance structure of $\mathcal{Z}(t)$, since $\mathbb{E}[\mathcal{D}(s)\mathcal{D}(t)] = \mathbb{E}[(N_1(s) - \mu s)(N_1(t) - \mu t)]$. Let $N^e(t)$ denote an equilibrium renewal process with renewal distribution S, and $N^o(t)$ denote an ordinary renewal process with renewal distribution S. Let $C_1 \triangleq \mu c_S^2$, $C_2 \triangleq \frac{4}{3}\mu^3\mathbb{E}[S^3] + \frac{1}{4}\mu^4(\mathbb{E}[S^2])^2$, and $C_3 \triangleq \mu^3\mathbb{E}[S^2]$. Also, let $f(t) \triangleq V[N^e(t)] - C_1 t$. Then

Lemma 39. *$f(0) = 0$, and $\sup_{t\geq 0}|f(t)| \leq C_2$. $f(t)$ is Lipschitz, namely $|f(t+h) - f(t)| \leq C_3 h$ for all $t, h \geq 0$. Also, $\mathbb{E}[(N^e(s) - \mu s)(N^e(t) - \mu t)] = C_1 s + \frac{1}{2}(f(s) + f(t) - f(t-s))$ for all $s, t \geq 0$.*

Proof. That $f(0) = 0$ is trivial. That $\sup_{t\geq 0}|f(t)| \leq C_2$ follows from [29] Equation 1.15. We now prove that $|f(t+h) - f(t)| \leq C_3 h$ for all $t, h \geq 0$. Noting that $\mathbb{E}[N^o(t)]$ is monotone and bounded on compact sets and thus integrable, it follows from [29] Equation 1.4 that

$$f(t) = 2\mu \int_0^t \left(\left(\mathbb{E}[N^o(s)] + 1 - \mu s\right) - \frac{1}{2}(1 + c_S^2) \right) ds. \tag{5.25}$$

It is proven in [73] that for all $s \geq 0$, one has $0 \leq \mathbb{E}[N^o(s)] + 1 - \mu s \leq \mu^2\mathbb{E}[S^2]$, and it follows that

$$\left| \left(\mathbb{E}[N^o(s)] + 1 - \mu s\right) - \frac{1}{2}(1 + c_S^2) \right| \leq \frac{1}{2}\mu^2\mathbb{E}[S^2].$$

Combining with (5.25) completes the proof.

We now prove the final assertion of the lemma. $\mathbb{E}[(N^e(s)-\mu s)(N^e(t)-\mu t)]$ equals

$$\begin{aligned} & \mathbb{E}[N^e(s)N^e(t)] - \mu^2 st \\ = & -\frac{1}{2}\Big(\mathbb{E}[\big(N^e(t) - N^e(s)\big)^2] - \mathbb{E}[\big(N^e(s)\big)^2] - \mathbb{E}[\big(N^e(t)\big)^2]\Big) - \mu^2 st \\ = & \frac{1}{2}\big(V[N^e(t)] + V[N^e(s)] - V[N^e(t-s)]\big) \text{ by stationary increments,} \end{aligned}$$

and the assertion then follows from definitions, completing the proof of the lemma. □

We conclude this subsection by showing that the covariance of the number of renewals at different times of an equilibrium renewal process is always non-negative.

Lemma 40. $\mathbb{E}[N^e(s)N^e(t)] - \mu^2 st \geq 0$ *for all* $s, t \geq 0$.

Proof. The proof is deferred to the appendix. □

5.6.4 Bounding the covariance of $\mathcal{Z}(t)$

In this subsection, we compare the covariance of $\mathcal{Z}(t)$ to that of a combination of a Brownian motion and an O-U process. Let $e \triangleq \exp(1)$, and $\epsilon_0 \triangleq \big(2(e-2)\big)^{-1}$. It can be easily verified that $\exp(-\epsilon_0) < \frac{1}{2}$. Let $C_4 \triangleq \mu c_A^2 + C_1$, and $M \triangleq \frac{8(C_4+C_2+C_3)}{C_4\left(\frac{1}{2}-\exp(-\epsilon_0)\right)}$. Let $\mathcal{U}_1^M(t)$ denote a realization of the process $\mathcal{U}^M(t)$, independent of $\{B_i^b(t), b \in \mathbb{R}, i \geq 1\}$. Let us define a new Gaussian process $\mathcal{W}(s)$ on $[M, \infty)$. Note that $MC_4 + f(s) > 0$ for all $s \geq 0$ by the construction of M. We define

$$\mathcal{W}(s) \triangleq C_4^{\frac{1}{2}}(1 - \frac{M}{s})^{\frac{1}{2}}\mathcal{B}^0(s) + \big(MC_4 + f(s)\big)^{\frac{1}{2}}\mathcal{U}^M(s), s \geq M + 1.$$

That $\mathcal{W}(t)$ is a Gaussian process follows from the fact that sums of Gaussian processes are Gaussian processes, and a Gaussian process multiplied by a deterministic function of time is a Gaussian process. That $\mathcal{W}(t)$ is continuous follows from the continuity of $\mathcal{B}_1^0(t)$ and $\mathcal{U}_1^M(t)$, combined with the fact that $f(s)$ is Lipschitz by Lemma 39. We now prove that

Theorem 40. *For all* $s \geq M+1$ *and* $t \geq s$, $V[\mathcal{W}(s)] = V[\mathcal{Z}(s)]$, *and* $V[\mathcal{Z}(s), \mathcal{Z}(t)] \geq V[\mathcal{W}(s), \mathcal{W}(t)]$.

Proof. The proof is deferred to the appendix. □

We conclude this subsection by showing that, in addition to Theorem 40, $\mathcal{Z}(t)$ satisfies

Theorem 41. $V[\mathcal{Z}(s), \mathcal{Z}(t)] \geq 0$ *for all* $s, t \geq 0$.

Proof. Note that

$$\begin{aligned} V[\mathcal{Z}(s), \mathcal{Z}(t)] &= V[\mathcal{A}(s), \mathcal{A}(t)] + V[\mathcal{D}(s), \mathcal{D}(t)] \\ &= \mu c_A^2 s + \mathbb{E}[N^e(s) N^e(t)] - \mu^2 st \quad \geq \quad 0 \quad \text{by Lemma 40.} \end{aligned}$$

□

5.6.5 Proof of Theorem 25

In this subsection, we complete the proof of Theorem 25. It will first be useful to restate Theorem 28 as a lower bound on the probability that $\big(Q_B^n(\infty) - n\big)n^{-\frac{1}{2}} \leq x$, which follows immediately from Theorem 28 by taking complements.

Corollary 13. *For all* $B > 0$ *and* $x \in \mathbb{R}$, $\liminf_{n\to\infty} \mathbb{P}\Big(\big(Q_B^n(\infty) - n\big)n^{-\frac{1}{2}} \leq x\Big)$ *is at least*

$$\sup_{\substack{\delta \geq 0 \\ \eta \geq 0}} \mathbb{P}\Bigg(\max\Bigg(\sup_{0 \leq t \leq \delta} \big(\mathcal{Z}(t) + (\eta - B)\mu t\big) \ ,$$

$$\sup_{t \geq \delta} \big(\mathcal{Z}(t) - B\mu t\big) \ + \ \eta\mu\delta\Bigg) \ \leq \ x + \eta\mathbb{P}\big(R(X) \leq \delta\big)\Bigg).$$

We now conclude the proof of Theorem 25.

Proof of Theorem 25. Suppose $B \leq 1$. Recall that $C_4 = \mu(c_A^2 + c_S^2)$, and $M = \frac{8(C_4+C_2+C_3)}{C_4\left(\frac{1}{2}-\exp(-\epsilon_0)\right)}$, and thus C_4 and M are both finite, strictly positive constants depending only on the distributions of A and S. Since $\mathbb{E}[S^2] < \infty$ implies $\mathbb{E}[R(S)] < \infty$, we may define $\delta_0 \overset{\Delta}{=} 2M + 1 + 2\mathbb{E}[R(S)]$, and note that δ_0 is a finite, strictly positive constant depending only on the distributions of A and S. Since $\mathcal{D}(t)$ is continuous, $\sup_{0\leq t\leq\delta_0}\mathcal{D}(t)$ is a.s. finite. Thus we may select a constant $\eta_0 \in (1,\infty)$, depending only on the distribution of S, s.t. $\mathbb{P}\Bigg(\sup_{0\leq t\leq\delta_0}\mathcal{D}(t) \leq \frac{1}{4}\eta_0\Bigg) \geq \frac{1}{2}$. It follows from Theorem 41 and Corollary 12 that

$$\mathbb{P}\Bigg(\max\Bigg(\sup_{0\leq t\leq\delta_0}\big(\mathcal{Z}(t)+(\eta_0-B)\mu t\big)\ ,\ \sup_{t\geq\delta_0}\big(\mathcal{Z}(t)-B\mu t\big)+\eta_0\mu\delta_0\Bigg)\ \leq\ \eta_0\mathbb{P}\big(R(X)\leq\delta_0\big)\Bigg) \tag{5.26}$$

$$\geq\ \mathbb{P}\Bigg(\sup_{0\leq t\leq\delta_0}\big(\mathcal{Z}(t)+(\eta_0-B)\mu t\big)\ \leq\ \eta_0\mathbb{P}\big(R(X)\leq\delta_0\big)\Bigg) \tag{5.27}$$

$$\times\ \ \mathbb{P}\Bigg(\sup_{t\geq\delta_0}\big(\mathcal{Z}(t)-B\mu t\big)\ +\ \eta_0\mu\delta_0\ \leq\ \eta_0\mathbb{P}\big(R(X)\leq\delta_0\big)\Bigg). \tag{5.28}$$

We now bound (5.27). Note that since $\mathcal{D}(t)$ has the same distribution as $-\mathcal{D}(t)$, it follows from the independence of $\mathcal{A}(t)$ and $\mathcal{D}(t)$, and the fact that by construction δ_0 is at least the median of $R(S)$, that (5.27) is at least

$$\mathbb{P}\Bigg(\sup_{0\leq t\leq\delta_0}\big(\mathcal{A}(t)+(\eta_0-B)\mu t\big)\ \leq\ \frac{1}{4}\eta_0\Bigg) \tag{5.29}$$

$$\times\ \ \mathbb{P}\Bigg(\sup_{0\leq t\leq\delta_0}\mathcal{D}(t)\leq\ \frac{1}{4}\eta_0\Bigg). \tag{5.30}$$

It is a straightforward exercise to demonstrate that (5.29) is some strictly positive probability ϵ_1, depending only on the distributions of A and S. By the construction of η_0, (5.30) is at least $\frac{1}{2}$. Combining the above, we find that (5.27) is at least

$$\frac{1}{2}\epsilon_1. \tag{5.31}$$

We now bound (5.28). Since Corollary 5.i of Chapter 4 implies that $\sup_{t\geq 0}\big(\mathcal{Z}(t) - B\mu t\big)$ is a.s. finite, it follows from Theorem 40, Corollary 11, and the fact that $\delta_0 \geq M + 1$ that (5.28) is at least

$$\mathbb{P}\Bigg(\sup_{t\geq\delta_0}\bigg(C_4^{\frac{1}{2}}(1-\frac{M}{t})^{\frac{1}{2}}\mathcal{B}_1^0(t) + \big(MC_4 + f(t)\big)^{\frac{1}{2}}\mathcal{U}_1^M(t) - B\mu t\bigg) \ \leq\ -\eta_0\mu\delta_0\Bigg). \tag{5.32}$$

Note that $\sup_{t\geq 0}|\big(MC_4 + f(t)\big)^{\frac{1}{2}}| \leq (MC_4 + C_2)^{\frac{1}{2}}$ by Lemma 39. Thus it follows from Theorem 35 and the independence of $\mathcal{B}_1^0(t)$ and $\mathcal{U}_1^M(t)$ that we may construct an a.s. finite r.v. Z_1, whose distribution depends only on A and S, on the same probability space as $\mathcal{B}_1^0(t)$, s.t. Z_1 and $\mathcal{B}_1^0(t)$ are independent, and (5.32) is at least

$$\mathbb{P}\Bigg(\sup_{t\geq\delta_0}\bigg(C_4^{\frac{1}{2}}(1-\frac{M}{t})^{\frac{1}{2}}\mathcal{B}_1^0(t) + Z_1 + 2\log^{\frac{1}{2}}(t+1) - B\mu t\bigg) \ \leq\ -\eta_0\mu\delta\ \Bigg). \tag{5.33}$$

It is easily verified that $2\log^{\frac{1}{2}}(t+1) \leq t^{\frac{1}{4}} + 8$ for all $t \geq 0$. Since Z_1 is a.s. finite, it thus follows from (5.33) that there exists a finite constant $H_1 \in (1, \infty)$, depending only on the distributions of A and S, s.t. (5.32), and thus (5.28), is at least

$$\frac{1}{2}\mathbb{P}\Bigg(\sup_{t\geq\delta_0}\bigg(C_4^{\frac{1}{2}}(1-\frac{M}{t})^{\frac{1}{2}}\mathcal{B}_1^0(t) + t^{\frac{1}{4}} - B\mu t\bigg) \ \leq\ -\big(\eta_0\mu\delta_0 + H_1\big)\ \Bigg). \tag{5.34}$$

Let $f_1(t) \triangleq C_4^{\frac{1}{2}}(1 - \frac{M}{t})^{\frac{1}{2}}$. Note that the event $\big\{f_1(t)\mathcal{B}_1^0(t) + t^{\frac{1}{4}} - B\mu t > -\big(\eta_0\mu\delta_0 + H_1\big)\big\}$ is equivalent to the event $\big\{\mathcal{B}_1^0(t) + t^{\frac{1}{4}}\big(f_1(t)\big)^{-1} - B\mu t\big(f_1(t)\big)^{-1} > -\big(\eta_0\mu\delta_0 +$

$H_1)\big(f_1(t)\big)^{-1}\}$. Since $t \geq 2M$ implies $\frac{1}{2}C_4^{\frac{1}{2}} \leq f_1(t) \leq C_4^{\frac{1}{2}}$, it follows that for $t \geq \delta_0$, the event

$$\{\mathcal{B}_1^0(t) + t^{\frac{1}{4}}\big(f_1(t)\big)^{-1} - B\mu t\big(f_1(t)\big)^{-1} > -\big(\eta_0\mu\delta_0 + H_1\big)\big(f_1(t)\big)^{-1}\}$$

implies the event

$$\{\mathcal{B}_1^0(t) + t^{\frac{1}{4}}\big(\frac{1}{2}C_4^{\frac{1}{2}}\big)^{-1} - B\mu t\big(C_4^{\frac{1}{2}}\big)^{-1} > -\big(\eta_0\mu\delta_0 + H_1\big)\big(\frac{1}{2}C_4^{\frac{1}{2}}\big)^{-1}\}.$$

Let $C_5 \triangleq 2(\eta_0\mu\delta_0 + H_1 + 1)(1 + C_4^{-\frac{1}{2}} + \mu^{-1}C_4^{\frac{1}{2}})$, and note that C_5 is a finite constant depending only on the distributions of A and S. Combining the above with (5.34), it follows that (5.28) is at least

$$\frac{1}{2}\mathbb{P}\Bigg(\sup_{t \geq \delta_0}\Big(\mathcal{B}_1^0(t) + C_5 t^{\frac{1}{4}} - C_5^{-1}Bt\Big) \leq -C_5\Bigg). \tag{5.35}$$

By Theorem 38, i.e. the law of the iterated logarithm for the three-dimensional Bessel process, there exists an a.s. finite non-negative r.v. Z_2, whose distribution depends only on the distributions of A and S, s.t. $\mathcal{S}_1^1(t) \geq t^{\frac{3}{8}} - Z_2$ for all $t \geq 0$. Let C_6 denote the median of Z_2, and note that C_6 is a finite non-negative constant depending only on the distributions of A and S. Let $H_2 \triangleq C_5(1 + \delta_0^{\frac{1}{4}}) + C_6 + C_5^3 - 1$, and note that H_2 is a finite non-negative constant depending only on the distributions of A and S. Then (5.35) is at least

$$\frac{1}{2}\mathbb{P}\big(\mathcal{B}_1^0(\delta_0) \leq -H_2\big)\mathbb{P}\Bigg(\sup_{t \geq 0}\Big(\mathcal{B}_1^0(t) + C_5(t+\delta_0)^{\frac{1}{4}} - C_5^{-1}B(t+\delta_0)\Big) \leq H_2 - C_5\Bigg). \tag{5.36}$$

Let $\epsilon_2 \triangleq \frac{1}{2}\mathbb{P}\big(\mathcal{B}_1^0(\delta_0) \leq -H_2\big)$, and note that $\epsilon_2 \in (0,1)$ depends only on the distributions of A and S. Then since $(x+y)^{\frac{1}{4}} \leq x^{\frac{1}{4}} + y^{\frac{1}{4}}$ for all $x, y \geq 0$, and $\delta_0 \geq 0$, it follows from (5.36) that (5.28) is at least

$$\epsilon_2 \mathbb{P}\left(\sup_{t \geq 0} \left(\mathcal{B}_1^0(t) + C_5 t^{\frac{1}{4}} - C_5^{-1} B t \right) \leq H_2 - C_5 (1 + \delta_0^{\frac{1}{4}}) \right). \tag{5.37}$$

Since $\mathcal{B}_1^0(t)$ has the same distribution as $-\mathcal{B}_1^0(t)$, note that (5.37) equals

$$\epsilon_2 \mathbb{P}\left(\inf_{t \geq 0} \left(\mathcal{B}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} B t \right) \geq -\left(C_6 + C_5^3\right) \right). \tag{5.38}$$

Let $C_7 \triangleq -\left(C_6 + C_5^3\right)$, and note that C_7 is a finite, negative constant which depends only on the distributions of A and S. We now bound (5.38) from below by conditioning on an appropriate event related to hitting times, and then applying Theorem 37, i.e. the relationship between Brownian motion with appropriately conditioned hitting times and the three-dimensional Bessel process. It follows from Theorem 39, applied with $b = 1$ and $\epsilon = \frac{1}{8}$, that there exist absolute finite constants $\delta, T > 0$ s.t. for all $x \geq T$, one has

$$\mathbb{P}\left(\tau_{\mathcal{S}^1}^{xB^{-1}} < \delta x^2 B^{-1}\right) \leq \frac{1}{8}.$$

Let $H_3 \triangleq T C_5 + 2^{\frac{2}{3}} C_5^{\frac{4}{3}} \delta^{-\frac{1}{2}}$, and $H_4 \triangleq \left(2 C_5^2\right)^{\frac{4}{3}}$, and note that both H_3 and H_4 are finite, strictly positive constants depending only on the distributions of A and S, s.t. $\mathbb{P}\left(\tau_{\mathcal{S}_1^1}^{H_3 B^{-1}} < H_4 B^{-2} \right) \leq \frac{1}{8}$. Note that (5.38), and thus (5.28), is at least

$$\epsilon_2 \mathbb{P}\left(\tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} < \tau_{\mathcal{B}_1^1}^{0} \right) \tag{5.39}$$

$$\times \mathbb{P}\left(\tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} \geq H_4 B^{-2}, \inf_{0 \leq t \leq \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}} \left(\mathcal{B}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} B t \right) \geq C_7, \right. \tag{5.40}$$

$$\left. \inf_{t \geq \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}} \left(\mathcal{B}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} B t \right) \geq C_7 \middle| \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} < \tau_{\mathcal{B}_1^1}^{0} \right).$$

We now bound (5.40). By construction, $t \geq H_4 B^{-2}$ implies $C_5 t^{\frac{1}{4}} \leq \frac{1}{2} C_5^{-1} Bt$. It follows that (5.40) is at least

$$\mathbb{P}\Bigg(\tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} \geq H_4 B^{-2}, \quad \inf_{0 \leq t \leq \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}} \Big(\mathcal{B}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} Bt \Big) \geq C_7, \tag{5.41}$$

$$\inf_{t \geq \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}} \Big(\mathcal{B}_1^1(t) + \frac{1}{2} C_5^{-1} Bt \Big) \geq C_7 \Big| \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} < \tau_{\mathcal{B}_1^1}^{0} \Bigg).$$

It then follows from a union bound that (5.40) is at least

$$\mathbb{P}\Bigg(\inf_{0 \leq t \leq \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}} \Big(\mathcal{B}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} Bt \Big) \geq C_7, \tag{5.42}$$

$$\inf_{t \geq \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}} \Big(\mathcal{B}_1^1(t) + \frac{1}{2} C_5^{-1} Bt \Big) \geq C_7 \Big| \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} < \tau_{\mathcal{B}_1^1}^{0} \Bigg)$$

$$-\mathbb{P}\Bigg(\tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} < H_4 B^{-2} \Big| \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} < \tau_{\mathcal{B}_1^1}^{0} \Bigg). \tag{5.43}$$

We now bound (5.42). It follows from the strong Markov property of Brownian motion, and the fact that $C_5^{-1} B\big(t + \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}\big) \geq C_5^{-1} Bt$, that (5.42) is at least

$$\mathbb{P}\Bigg(\inf_{t \geq 0} \Big(H_3 B^{-1} + \mathcal{B}_1^0(t) + \frac{1}{2} C_5^{-1} Bt \Big) \geq C_7 \Bigg) \tag{5.44}$$

$$\times \mathbb{P}\Bigg(\inf_{0 \leq t \leq \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}}} \Big(\mathcal{B}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} Bt \Big) \geq C_7 \Big| \tau_{\mathcal{B}_1^1}^{H_3 B^{-1}} < \tau_{\mathcal{B}_1^1}^{0} \Bigg). \tag{5.45}$$

We now bound (5.45). It follows from Theorem 37 that (5.45) equals

$$\mathbb{P}\Bigg(\inf_{0 \leq t \leq \tau_{\mathcal{S}_1^1}^{H_3 B^{-1}}} \Big(\mathcal{S}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1} Bt \Big) \geq C_7 \Bigg), \tag{5.46}$$

which is at least

$$\mathbb{P}\Bigg(\inf_{t\geq 0}\Big(\mathcal{S}_1^1(t) - C_5 t^{\frac{1}{4}} + C_5^{-1}Bt\Big) \geq C_7\Bigg). \tag{5.47}$$

By Theorem 38, i.e. the law of the iterated logarithm for the three-dimensional Bessel process, and the definition of C_6, it follows that (5.47) is at least

$$\frac{1}{2}\mathbb{P}\Bigg(\inf_{t\geq 0}\Big(t^{\frac{3}{8}} - C_5 t^{\frac{1}{4}} + C_5^{-1}Bt\Big) \geq C_6 + C_7\Bigg). \tag{5.48}$$

Since by construction $t^{\frac{3}{8}} \geq C_6 + C_7 + C_5 t^{\frac{1}{4}}$ for all $t \geq 0$, it follows that (5.45) is at least $\frac{1}{2}$. We now bound (5.44). Since $\mathcal{B}_1^0(t)$ has the same distribution as $-\mathcal{B}_1^0(t)$, it follows from Theorem 33.iii that (5.44) equals

$$\mathbb{P}\Bigg(\sup_{t\geq 0}\Big(\mathcal{B}_1^0(t) - \frac{1}{2}C_5^{-1}Bt\Big) \leq H_3 B^{-1} - C_7\Bigg) = 1 - \exp\big(-H_3 C_5^{-1} + BC_7 C_5^{-1}\big). \tag{5.49}$$

Combining our bounds for (5.44) and (5.45) with the fact that by construction $\exp\big(-H_3 C_5^{-1} + BC_7 C_5^{-1}\big) \leq \frac{1}{2}$, it follows that (5.42) is at least $\frac{1}{4}$. We now bound (5.43). It follows from Theorem 37 that (5.43) equals

$$-\mathbb{P}\Bigg(\tau_{\mathcal{S}_1^1}^{H_3 B^{-1}} < H_4 B^{-2}\Bigg), \tag{5.50}$$

which by construction is at least $-\frac{1}{8}$. Combining our bounds for (5.42) and (5.43), it follows that (5.40) is at least $\frac{1}{8}$. It follows from Theorem 33.ii and the symmetries of Brownian motion that (5.39) equals $\epsilon_2 H_3^{-1} B$. Combining our bounds for (5.40) and (5.39), it follows that (5.28) is at least $\frac{1}{8}\epsilon_2 H_3^{-1} B$. Combining with (5.31), our bound for (5.27), it follows that (5.26) is at least $\frac{1}{16}\epsilon_1\epsilon_2 H_3^{-1} B$. Since ϵ_1, ϵ_2, and H_3^{-1} are all strictly positive constants depending only on the distributions of A and S, and not B, this concludes the proof of the theorem. □

5.7 Proof of Large Deviations Result

In this section we complete the proof of Theorem 26.

Proof of Theorem 26. Let us fix some $B > 0$, and $x < -1$. Note that it follows from the proof of Theorem 12 given in Chapter 4 that $\sup_{t\geq 0}\big(\mathcal{D}(t) - \frac{B}{2}\mu t\big)$ is a.s. finite. Thus we may select a constant $c \in (8,\infty)$, depending only on the distribution of S, s.t.

$$\mathbb{P}\bigg(\sup_{t\geq 0}\big(\mathcal{D}(t) - \frac{B}{2}\mu t\big) \leq \frac{c}{8}\bigg) \geq \frac{1}{2}. \tag{5.51}$$

Let $\delta_0 \triangleq 2\big(1 + (16\mu)^{-1}\big)\mathbb{E}[R(S)]$. Then $\mathbb{P}\big(R(X) \leq \delta_0\big) \geq \frac{1}{2}$, and it follows from Corollary 13, with $\delta = \delta_0$ and $\eta = c|x|$, that

$$\liminf_{n\to\infty} \mathbb{P}\bigg(\big(Q^n_B(\infty) - n\big)n^{-\frac{1}{2}} \leq x\bigg) \tag{5.52}$$

is at least

$$\mathbb{P}\Bigg(\max\bigg(\sup_{0\leq t\leq\delta_0}\big(\mathcal{A}(t) - \mathcal{D}(t) + (c|x| - B)\mu t\big),$$

$$\sup_{t\geq\delta_0}\big(\mathcal{A}(t) - \mathcal{D}(t) - B\mu t\big) \ + \ c|x|\mu\delta_0\bigg) \leq (\frac{c}{2} - 1)|x|\Bigg)$$

$$\geq \ \mathbb{P}\Bigg(\max\bigg(\sup_{0\leq t\leq\delta_0}\big(\mathcal{A}(t) + (c|x| - \frac{B}{2})\mu t\big) + \sup_{0\leq t\leq\delta_0}\big(-\mathcal{D}(t) - \frac{B}{2}\mu t\big),$$

$$\sup_{t\geq\delta_0}\big(\mathcal{A}(t) - \frac{B}{2}\mu t + c|x|\mu\delta_0\big) \ + \ \sup_{t\geq\delta_0}\big(-\mathcal{D}(t) - \frac{B}{2}\mu t\big)\bigg) \ \leq \ (\frac{c}{2} - 1)|x|\Bigg)$$

$$\geq \mathbb{P}\Bigg(\max\bigg(\sup_{0\leq t\leq\delta_0}\big(\mathcal{A}(t) + (c|x| - \frac{B}{2})\mu t\big) + \sup_{t\geq 0}\big(-\mathcal{D}(t) - \frac{B}{2}\mu t\big),$$

$$\sup_{t\geq\delta_0}\big(\mathcal{A}(t) - \frac{B}{2}\mu t + c|x|\mu\delta_0\big) \ + \ \sup_{t\geq 0}\big(-\mathcal{D}(t) - \frac{B}{2}\mu t\big)\bigg) \ \leq \ (\frac{c}{2} - 1)|x|\Bigg),$$

which is at least

$$\mathbb{P}\Bigg(\sup_{t\geq 0} \big(-\mathcal{D}(t) - \frac{B}{2}\mu t\big) + \max\Bigg(\sup_{0\leq t\leq \delta_0} \big(\mathcal{A}(t) + (c|x| - \frac{B}{2})\mu t\big) \ , \tag{5.53}$$

$$\sup_{t\geq \delta_0} \big(\mathcal{A}(t) - \frac{B}{2}\mu t + c|x|\mu\delta_0\big)\Bigg) \ \leq \ \frac{c}{4}|x| \Bigg),$$

since $c > 4$ implies $\frac{c}{2} - 1 > \frac{c}{4}$. It follows from (5.53), a union bound, and the definition of c that (5.52) is at least

$$\frac{1}{2}\mathbb{P}\Bigg(\max\Bigg(\sup_{0\leq t\leq \delta_0} \big(\mathcal{A}(t) + (c|x| - \frac{B}{2})\mu t\big) \ , \ \sup_{t\geq \delta_0} \big(\mathcal{A}(t) - \frac{B}{2}\mu t + c|x|\mu\delta_0\big)\Bigg) \ \leq 0 \Bigg). \tag{5.54}$$

Although an exact analysis of expression (5.54) follows from well-known results about Brownian motion (see e.g. [98],[8],[11]), a simpler analysis suffices for our purposes, which we include for completeness. Note that (5.54) is at least

$$\frac{1}{2}\mathbb{P}\Bigg(\max\Bigg(\sup_{0\leq t\leq \delta_0} \big(\mathcal{A}(t) + c|x|\mu t\big) \ , \ \sup_{t\geq \delta_0} \big(\mathcal{A}(t) - \frac{B}{2}\mu t + c|x|\mu\delta_0\big)\Bigg) \ \leq \ 1 \Bigg). \tag{5.55}$$

Also, since $c|x|\mu t \leq c|x|\mu\delta_0$ on $[0, \delta_0]$, the event

$$\Bigg\{ \sup_{0\leq t\leq (16\mu)^{-1}} \mathcal{A}(t) \leq 1, \mathcal{A}\big((16\mu)^{-1}\big) \leq -2c|x|\mu\delta_0 - 2, \tag{5.56}$$

$$\sup_{(16\mu)^{-1}\leq t\leq \delta_0} \Bigg(\mathcal{A}(t) - \mathcal{A}\big((16\mu)^{-1}\big)\Bigg) \leq 1\Bigg\}.$$

implies the event

$$\Bigg\{ \sup_{0\leq t\leq \delta_0} \big(\mathcal{A}(t) + c|x|\mu t\big) \leq 1, \mathcal{A}(\delta_0) \leq -c|x|\mu\delta_0 \Bigg\}. \tag{5.57}$$

Note that the probability of the event

$$\Big\{ \sup_{0 \le t \le (16\mu)^{-1}} \mathcal{A}(t) \le 1, \mathcal{A}\big((16\mu)^{-1}\big) \le -2c|x|\mu\delta_0 - 1 \Big\}$$

is at least the probability of the event

$$\Big\{ \Big(\sup_{0 \le t \le (16\mu)^{-1}} \mathcal{A}(t) \Big) \Big(\sup_{0 \le t \le (16\mu)^{-1}} \mathcal{A}(t) - \mathcal{A}\big((16\mu)^{-1}\big) \Big) \le 1, \mathcal{A}\big((16\mu)^{-1}\big) \le -2c|x|\mu\delta_0 - 1 \Big\}.$$

It thus follows from Theorem 34 and the strong Markov property of Brownian motion that the probability of event (5.56), and thus (5.57), is at least

$$\mathbb{P}\Big(\frac{1}{2}(16\mu)^{-1}e_1 \le 1 \Big) \mathbb{P}\Big(\mathcal{A}\big((16\mu)^{-1}\big) \le -2c|x|\mu\delta_0 - 1 \Big) \mathbb{P}\Big(\sup_{0 \le t \le \delta_0} \mathcal{A}(t) \le 1 \Big). \quad (5.58)$$

It is straightforward to demonstrate that there exists a finite strictly positive constant c_1, depending only on the distributions of A and S, s.t. (5.58) is at least $\exp(-c_1 x^2)$. Note that by the strong Markov property of Brownian motion, conditional on event (5.57), the probability of the event

$$\Big\{ \sup_{t \ge 0} \big(\mathcal{A}(t) - \frac{B}{2}\mu t\big) \le 1 \Big\},$$

which is just some finite strictly positive constant c_2, depending only on B and the distributions of A and S. Combining the above, we find that (5.55) is at least $\frac{1}{2}c_2 \exp(-c_1 x^2)$ for all $x \le -1$, completing the proof. □

5.8 Comparison to Other Bounds From the Literature

In this section we compare our results to known results for the $GI/M/n$ and $M/GI/n$ queues, showing that our main results are tight, in an appropriate sense. Recall that G denotes a standard normal r.v. We also define $\alpha(B) \triangleq \lim_{n\to\infty} \mathbb{P}(Q^n_B(\infty) \geq n)$, i.e. the limiting steady-state probability of delay, when this limit exists.

5.8.1 $GI/M/n$ queue

In [52], Halfin and Whitt proved that

Theorem 42. *If Q^n_B is a $GI/M/n$ queue, then*

$$\alpha(B) = \bigg(1 + (2\pi)^{\frac{1}{2}} \frac{2B}{1+c_A^2} \mathbb{P}(G \leq \frac{2B}{1+c_A^2}) \exp\big(\frac{1}{2}(\frac{2B}{1+c_A^2})^2\big)\bigg)^{-1}. \tag{5.59}$$

For all $x \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(Q^n_B(\infty) \geq n + xn^{\frac{1}{2}}) = \alpha(B)\exp(-\frac{2B}{1+c_A^2}x); \tag{5.60}$$

and

$$\lim_{n\to\infty} \mathbb{P}(Q^n_B(\infty) \leq n - xn^{\frac{1}{2}}) = \big(1 - \alpha(B)\big)\frac{\mathbb{P}(G \leq B - x)}{\mathbb{P}(G \leq B)}. \tag{5.61}$$

It follows from Theorem 42 and elementary asymptotics that

Corollary 14. *If Q^n_B is a $GI/M/n$ queue, then*

$$\lim_{B\to\infty} B^{-2}\log\alpha(B) = -2(1+c_A^2)^{-2} > -\infty,$$

$$\lim_{B\to 0} B^{-1}\big(1-\alpha(B)\big) = (2\pi)^{\frac{1}{2}}(1+c_A^2)^{-1} < \infty,$$

and

$$\lim_{x\to\infty} x^{-2}\log\left(\lim_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)n^{-\frac{1}{2}}<-x\Big)\right)=-\frac{1}{2}<0.$$

Thus in this case, we conclude that all of our main results correctly capture the correct qualitative scaling of the relevant quantities. Namely, the limiting steady-state probability of delay scales like $\exp\big(-\epsilon_1 B^2\big)$ as $B\to\infty$ for some $\epsilon_1>0$; the limiting steady-state probability that a job does not have to wait for service scales like $\epsilon_2 B$ as $B\to 0$ for some $\epsilon_2>0$; and the tail of the limiting steady-state number of idle servers scales like $\exp\big(-\epsilon_3 x^2\big)$ as $x\to\infty$ for some $\epsilon_3>0$.

5.8.2 $M/GI/n$ queue

Suppose that Q^n_B is an $M/GI/n$ queue. Let $Z_{n,B}$ denote a Poisson r.v. with mean $\lambda_{n,B}$. Then it follows from a naive infinite-server lower bound, and the well-known properties of the steady-state infinite server queue (see [101]), that for all $x\in\mathbb{R}^+$, $\mathbb{P}\big(Q^n_B(\infty)\leq n-xn^{\frac{1}{2}}\big)\leq\mathbb{P}\big(Z_{n,B}\leq n-xn^{\frac{1}{2}}\big)$. It follows from the Central Limit Theorem that for all $x\in\mathbb{R}^+$,

$$\lim_{n\to\infty}\mathbb{P}(Z_{n,B}\leq n-xn^{\frac{1}{2}})=\mathbb{P}\big(G\leq B-x\big).$$

It follows that

$$\liminf_{B\to\infty} B^{-2}\log\left(\liminf_{n\to\infty}\mathbb{P}\big(Q^n_B(\infty)\geq n\big)\right)\geq-\frac{1}{2}>-\infty,$$

and

$$\limsup_{x\to\infty} x^{-2}\log\left(\limsup_{n\to\infty}\mathbb{P}\Big(\big(Q^n(\infty)-n\big)n^{-\frac{1}{2}}<-x\Big)\right)\leq-\frac{1}{2}<0,$$

showing that in this setting Theorem 24 and Theorem 26 are again tight in the same sense as before. Interestingly, the infinite server lower-bound does not seem to yield any information about the tightness of Theorem 25, since $\mathbb{P}(Z_{n,B} \leq n) \geq \frac{1}{2}$ for all $B \geq 0$.

We note that the results of [60] for the case of deterministic processing times yield similar confirmation of our results, but we do not pursue that here.

5.9 Conclusion and Open Questions

In this chapter, we derived the first qualitative insights into the steady-state probability of delay in the H-W regime for generally distributed processing times. In particular, we derived bounds for the asymptotics of how the steady-state probability of delay scales as $B \rightarrow 0$ and $B \rightarrow \infty$, and found these bounds to be tight, in an appropriate sense, by comparing to known results for special cases. We also revisited the question of large deviations for the steady-state number of idle servers, and proved that this r.v. has a Gaussian-like tail.

Our main proof technique was the derivation of new and simple bounds for the FCFS $GI/GI/n$ queue. Our bounds are of a structural nature, hold for all n and all times $t \geq 0$, and have intuitive closed-form representations as the suprema of certain natural processes which converge weakly to Gaussian processes. Our bounds built on the techniques of Chapter 4, by allowing one to simultaneously keep all servers busy and alter the number of servers.

This work leaves many interesting directions for future research. In particular, it would be very interesting to derive better bounds for the steady-state probability of delay. Furthermore, it is an open challenge to identify the exact scaling behavior of the probability of delay, as $B \rightarrow \infty$ and $B \rightarrow 0$, for processing times which are neither Markovian [52] nor deterministic [60].

5.10 Appendix

5.10.1 Proof of Theorem 41

In this subsection we complete the proof of Theorem 41.

Proof of Theorem 41. Let $\{X_k, k \geq 1\}$ denote the ordered sequence of renewal intervals in process $N^e(t)$. Then from definitions,

$$\mathbb{E}[N^e(s)N^e(t)] - \mu^2 st = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)I(\sum_{k=1}^{j} X_k \leq t)] - \mu^2 st. \tag{5.62}$$

Note that for $j \leq i$, one has that

$$\begin{aligned}\mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)I(\sum_{k=1}^{j} X_k \leq t)] &= \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)] &(5.63)\\ &\geq \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)]\mathbb{E}[I(\sum_{k=1}^{j} X_k \leq t)]. &(5.64)\end{aligned}$$

Alternatively, suppose $j \geq i+1$. Let $Y^1 \triangleq \sum_{k=1}^{i} X_k$, $Y^2 \triangleq \sum_{k=i+1}^{j} X_k$, and $Y^3 \triangleq t - Y^1$. Then Y^1 and Y^2 are independent, Y^2 and Y^3 are independent, and $\mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)I(\sum_{k=1}^{j} X_k \leq t)]$ equals

$$\begin{aligned}&\mathbb{E}[I(Y^1 \leq s)I(Y^1 + Y^2 \leq t)]\\ &= \mathbb{E}[I(Y^3 \geq t-s)I(Y^3 \geq Y^2)]\\ &= \mathbb{E}[I(Y^2 \geq t-s)I(Y^3 \geq Y^2) + I(Y^2 < t-s)I(Y^3 \geq t-s)]. \qquad (5.65)\end{aligned}$$

Let Y_a^3, Y_b^3 denote two r.v.s, each distributed as Y^3, where Y_a^3, Y_b^3, Y^2 are mutually independent. Then (5.65) equals

$$\begin{aligned}
&\mathbb{E}[I(Y^2 \geq t-s)I(Y_a^3 \geq Y^2) + I(Y^2 < t-s)I(Y_b^3 \geq t-s)] \\
&\quad \geq \mathbb{E}[I(Y^2 \geq t-s)I(Y_a^3 \geq Y^2)I(Y_b^3 \geq t-s)] \\
&\qquad + \mathbb{E}[I(Y^2 < t-s)I(Y_a^3 \geq Y^2)I(Y_b^3 \geq t-s)] \\
&\quad = \mathbb{E}[I(Y_a^3 \geq Y^2)I(Y_b^3 \geq t-s)] \\
&\quad = \mathbb{E}[I(Y^1 + Y^2 \leq t)]\mathbb{E}[I(Y^1 \leq s)]. \\
&\quad = \mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)]\mathbb{E}[I(\sum_{k=1}^{j} X_k \leq t)]. \qquad (5.66)
\end{aligned}$$

Combining (5.62) - (5.66), we find that

$$\begin{aligned}
\mathbb{E}[N^e(s)N^e(t)] - \mu^2 st &\geq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\mathbb{E}[I(\sum_{k=1}^{i} X_k \leq s)]\mathbb{E}[I(\sum_{k=1}^{j} X_k \leq t)] - \mu^2 st \\
&= \mathbb{E}[\sum_{i=1}^{\infty} I(\sum_{k=1}^{i} X_k \leq s)]\mathbb{E}[\sum_{j=1}^{\infty} I(\sum_{k=1}^{j} X_k \leq t)] - \mu^2 st \\
&= \mathbb{E}[N^e(s)]\mathbb{E}[N^e(t)] - \mu^2 st \quad = \quad 0,
\end{aligned}$$

completing the proof. □

5.10.2 Proof of Theorem 39

In this subsection we complete the proof of Theorem 39.

Proof of Theorem 39. It follows from Theorem 36 that $U_b, \mathcal{B}_1^b(t), \mathcal{B}_2^b(t), \mathcal{S}_1^0(t)$, and $\mathcal{S}_1^b(t)$ can be constructed on the same probability space s.t.

$$\sup_{0 \leq t \leq \epsilon M^2} |\mathcal{S}_1^b(t)| \leq \sup_{0 \leq t \leq \tau_{\mathcal{B}_1^b}^{U_b}} |\mathcal{B}_1^b(t)| + \sup_{0 \leq t \leq \tau_{\mathcal{B}_1^b}^{U_b} + \tau_{\mathcal{B}_2^b}^{U_b}} |\mathcal{B}_2^b(t)| + \sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}_1^0(t) + b. \qquad (5.67)$$

Since $\tau^{U_b}_{\mathcal{B}^b_1} \leq \tau^0_{\mathcal{B}^b_1}$, and $\tau^{U_b}_{\mathcal{B}^b_2} \leq \tau^0_{\mathcal{B}^b_2}$, it follows from (5.67) and a union bound that

$$\mathbb{P}\big(\sup_{0 \leq t \leq \epsilon M^2} |\mathcal{S}^b_1(t)| > M\big)$$

is at most

$$\mathbb{P}\big(\tau^0_{\mathcal{B}^b_1} + \tau^0_{\mathcal{B}^b_2} > \epsilon M^2\big) + \mathbb{P}\big(\sup_{0 \leq t \leq \epsilon M^2} |\mathcal{B}^b_1(t)| > \frac{M}{3}\big) \tag{5.68}$$

$$+\mathbb{P}\big(\sup_{0 \leq t \leq \epsilon M^2} |\mathcal{B}^b_2(t)| > \frac{M}{3}\big) + \mathbb{P}\big(\sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}^0_1(t) + b > \frac{M}{3}\big).$$

Since $\mathcal{B}^b_1(t)$ and $\mathcal{B}^b_2(t)$ have the same distribution, it follows from (5.68), the triangle inequality, a union bound, and the fact that $b < \frac{M}{6}$ that $\mathbb{P}\big(\sup_{0 \leq t \leq \epsilon M^2} |\mathcal{S}^b_1(t)| > M\big)$ is at most

$$2\mathbb{P}\big(\tau^0_{\mathcal{B}^b_1} > \frac{\epsilon}{2} M^2\big) \tag{5.69}$$

$$+ \quad 2\mathbb{P}\big(\sup_{0 \leq t \leq \epsilon M^2} |\mathcal{B}^0_1(t)| > \frac{M}{6}\big) \tag{5.70}$$

$$+ \quad \mathbb{P}\big(\sup_{0 \leq t \leq \epsilon M^2} \mathcal{S}^0_1(t) > \frac{M}{6}\big). \tag{5.71}$$

We now bound (5.69). It follows from the symmetries of Brownian motion that $\tau^0_{\mathcal{B}^b_1}$ has the same distribution as $\tau^b_{\mathcal{B}^0}$. Recall that G is a normally distributed r.v. with mean 0 and variance 1, and let $f_G(x)$ denote the density function of G. It follows from Theorem 33.i and a union bound that (5.69) is at most

$$\begin{aligned} 2\mathbb{P}(\tau^b_{\mathcal{B}^0} > \frac{\epsilon}{2}M^2) &= 2\mathbb{P}(\sup_{0\leq t\leq \frac{\epsilon}{2}M^2} \mathcal{B}^0(t) < b) \\ &= 2\Big(1 - \mathbb{P}(|G| > b(\frac{\epsilon}{2}M^2)^{-\frac{1}{2}})\Big) \\ &= 4\int_0^{b(\frac{\epsilon}{2}M^2)^{-\frac{1}{2}}} f_G(x)dx \\ &\leq (2\pi)^{-\frac{1}{2}}b(\frac{\epsilon}{2}M^2)^{-\frac{1}{2}} \text{ since } f_G(x) \leq (2\pi)^{-\frac{1}{2}} \text{ for all } x. \quad (5.72) \end{aligned}$$

Similarly, by the symmetries of Brownian motion, (5.70) is at most

$$\begin{aligned} 4\mathbb{P}(\sup_{0\leq t\leq \epsilon M^2} \mathcal{B}^0(t) > \frac{M}{6}) &= 8\mathbb{P}(G > \frac{M}{6}(\epsilon M^2)^{-\frac{1}{2}}) \\ &\leq 8\exp\big(\frac{1}{2} - \frac{1}{6}\epsilon^{-\frac{1}{2}}\big) \text{ by a Chernoff bound.} \quad (5.73) \end{aligned}$$

Since $\mathcal{S}^0_1(t)$ has the same distribution as $\Big(\sum_{i=1}^3 (\mathcal{B}^0_i(t))^2\Big)^{\frac{1}{2}}$, and $\Big(\sum_{i=1}^3 (\mathcal{B}^0_i(t))^2\Big)^{\frac{1}{2}} \leq \sum_{i=1}^3 |\mathcal{B}^0_i(t)|$, it follows from a union bound that (5.71) is at most

$$\begin{aligned} 3\mathbb{P}(\sup_{0\leq t\leq \epsilon M^2} |\mathcal{B}^0(t)| > \frac{M}{18}) &\leq 6\mathbb{P}(\sup_{0\leq t\leq \epsilon M^2} \mathcal{B}^0(t) > \frac{M}{18}) \\ &= 12\mathbb{P}(G > \frac{M}{18}(\epsilon M^2)^{-\frac{1}{2}}) \text{ by Lemma 33.i} \\ &\leq 12\exp(\frac{1}{2} - \frac{1}{18}\epsilon^{-\frac{1}{2}}) \text{ by a Chernoff bound.} (5.74) \end{aligned}$$

The theorem then follows from (5.72),(5.73), and (5.74). □

5.10.3 Proof of Theorem 40

In this subsection we complete the proof of Theorem 40. Let us fix some $s \geq M + 1$ and $t \geq s$. We begin by establishing several technical preliminaries.

Lemma 41. *(i)* $\left(1-\frac{M}{s}\right)^{\frac{1}{2}}\left(1-\frac{M}{t}\right)^{\frac{1}{2}}s \leq s-M+\frac{(t-s)M}{2t}$.

(ii) $\left(1-\frac{M}{s}\right)^{\frac{1}{2}}\left(1-\frac{M}{t}\right)^{\frac{1}{2}}s \leq s-\frac{M}{2}$.

(iii) $\left(MC_4+f(s)\right)^{\frac{1}{2}}\left(MC_4+f(t)\right)^{\frac{1}{2}} \leq MC_4+f(s)+\frac{C_3}{2}(t-s)$.

(iv) $\left(MC_4+f(s)\right)^{\frac{1}{2}}\left(MC_4+f(t)\right)^{\frac{1}{2}} \leq MC_4+C_2$.

(v) $V[\mathcal{Z}(s),\mathcal{Z}(t)] \geq C_4 s+f(s)-C_3(t-s)$.

(vi) $V[\mathcal{Z}(s),\mathcal{Z}(t)] \geq C_4 s-3C_2$.

Proof. (i) follows from the fact that

$$\begin{aligned}(1-\frac{M}{s})^{\frac{1}{2}}(1-\frac{M}{t})^{\frac{1}{2}}s &= (s-M)(\frac{1-\frac{M}{t}}{1-\frac{M}{s}})^{\frac{1}{2}}\\ &= (s-M)(1+\frac{(t-s)M}{st-tM})^{\frac{1}{2}}\\ &\leq (s-M)(1+\frac{(t-s)M}{2(s-M)t}).\\ &= s-M+\frac{(t-s)M}{2t}.\end{aligned}$$

(ii) follows from the fact that

$$\begin{aligned}(1-\frac{M}{s})^{\frac{1}{2}}(1-\frac{M}{t})^{\frac{1}{2}}s &\leq (1-\frac{M}{s})^{\frac{1}{2}}s\\ &\leq s-\frac{M}{2} \text{ since } (x+1)^{\frac{1}{2}} \leq 1+\frac{x}{2}.\end{aligned}$$

(iii) follows from the fact that

$$\begin{aligned}
(MC_4+f(s))^{\frac{1}{2}}(MC_4+f(t))^{\frac{1}{2}} &= (MC_4+f(s))(1+\frac{f(t)-f(s)}{MC_4+f(s)})^{\frac{1}{2}} \\
&\leq (MC_4+f(s))\left(1+\frac{|f(t)-f(s)|}{2(MC_4+f(s))}\right) \\
&\leq (MC_4+f(s))\left(1+\frac{C_3(t-s)}{2(MC_4+f(s))}\right) \\
&= MC_4+f(s)+\frac{C_3}{2}(t-s).
\end{aligned}$$

(iv) follows from the definition of C_2. (v) follows from the fact that

$$\begin{aligned}
V[\mathcal{Z}(s),\mathcal{Z}(t)] &= V[\mathcal{A}(s),\mathcal{A}(t)]+V[\mathcal{D}(s),\mathcal{D}(t)] \\
&= C_4s+\frac{1}{2}\big(f(s)+f(t)-f(t-s)\big) \text{ by Lemma 39} \\
&\geq C_4s+f(s)-\frac{|f(t)-f(s)|+|f(t-s)|}{2} \text{ by the triangle inequality} \\
&\geq C_4s+f(s)-C_3(t-s).
\end{aligned}$$

(vi) follows from the fact that

$$\begin{aligned}
V[\mathcal{Z}(s),\mathcal{Z}(t)] &= C_4s+\frac{1}{2}(f(s)+f(t)-f(t-s)) \\
&\geq C_4s-3C_2.
\end{aligned}$$

□

Lemma 42. *For all* $y\in[0,\epsilon_0]$, *one has* $\exp(-y)\leq 1-\frac{y}{2}$. *For all* $y\geq\epsilon_0$, *one has* $\frac{1}{2}-e^{-y}\geq\frac{1}{2}-\exp(-\epsilon_0)>0$.

Proof. It follows from a simple Taylor-series expansion that $|\exp(y)-(1+y)|\leq(e-2)y^2$ for all $y\in[-1,1]$. Noting that $|(e-2)y^2|\leq\frac{1}{2}|y|$ if $|y|\leq\epsilon_0$ completes the proof. □

We now complete the proof of Theorem 40.

Proof of Theorem 40. First, note that

$$\begin{aligned} V[\mathcal{W}(s)] &= V[C_4^{\frac{1}{2}}(1-\frac{M}{s})^{\frac{1}{2}}\mathcal{B}_1^0(s)] + V[(MC_4+f(s))^{\frac{1}{2}}\mathcal{U}_1^M(s)] \\ &= C_4(1-\frac{M}{s})s + MC_4 + f(s) \quad = \quad V[\mathcal{Z}(s)]. \end{aligned} \tag{5.75}$$

In general,

$$\begin{aligned} V[\mathcal{W}(s),\mathcal{W}(t)] &= V[C_4^{\frac{1}{2}}(1-\frac{M}{s})^{\frac{1}{2}}\mathcal{B}_1^0(s), C_4^{\frac{1}{2}}(1-\frac{M}{t})^{\frac{1}{2}}\mathcal{B}_1^0(t)] \\ &\quad +V[(MC_4+f(s))^{\frac{1}{2}}\mathcal{U}_1^M(s), (MC_4+f(t))^{\frac{1}{2}}\mathcal{U}_1^M(t)] \\ &= C_4(1-\frac{M}{s})^{\frac{1}{2}}(1-\frac{M}{t})^{\frac{1}{2}}s \qquad (5.76) \\ &\quad +(MC_4+f(s))^{\frac{1}{2}}(MC_4+f(t))^{\frac{1}{2}}\exp\big(-M(t-s)\big). \qquad (5.77) \end{aligned}$$

We now treat two cases. First, suppose $t-s \le \epsilon_0 M^{-1}$. Then (5.76) is at most $C_4\big(s - M + \frac{(t-s)M}{2t}\big)$ by (i), and (5.77) is at most

$$\begin{aligned} & \big(MC_4+f(s)+\frac{C_3}{2}(t-s)\big)\exp\big(-M(t-s)\big) \text{ by } (iii) \\ \le\ & \big(MC_4+f(s)+\frac{C_3}{2}(t-s)\big)\big(1-\frac{M}{2}(t-s)\big) \text{ by Lemma 42} \\ =\ & MC_4+f(s)+\frac{C_3}{2}(t-s)-\frac{M^2C_4}{2}(t-s)-\frac{Mf(s)}{2}(t-s)-\frac{MC_3}{4}(t-s)^2 \\ \le\ & MC_4+f(s)+\big(\frac{C_3}{2}+\frac{MC_2}{2}-\frac{M^2C_4}{2}\big)(t-s). \end{aligned} \tag{5.78}$$

Combining the above, we find that

$$\begin{aligned} V[\mathcal{W}(s),\mathcal{W}(t)] &\leq C_4 s - MC_4 + \frac{MC_4}{2t}(t-s) + MC_4 + f(s) \qquad (5.79)\\ &\quad +(\frac{C_3}{2} + \frac{MC_2}{2} - \frac{M^2C_4}{2})(t-s) \\ &\leq C_4 s + f(s) + (\frac{C_3}{2} + \frac{MC_2}{2} + \frac{MC_4}{2t} - \frac{M^2C_4}{2})(t-s) \\ &\leq C_4 s + f(s) + \Big(2M\big(C_4 + C_2 + C_3\big) - \frac{C_4}{2}M^2\Big)(t-s)(5.80) \end{aligned}$$

It follows from (v) and (5.80) that $V[\mathcal{Z}(s),\mathcal{Z}(t)] - V[\mathcal{W}(s),\mathcal{W}(t)]$ is at least

$$\begin{aligned} &\Big(\frac{C_4}{2}M^2 - 2M\big(C_4+C_2+C_3\big) - C_3\Big)(t-s) \\ &\geq M\Big(\frac{C_4}{2}(\frac{1}{2} - \exp(-\epsilon_0))M - 4\big(C_4+C_2+C_3\big)\Big)(t-s) \geq 0,(5.81) \end{aligned}$$

by the definition of M. Alternatively, suppose $t - s \geq \epsilon_0 M^{-1}$. Then (5.76) is at most $C_4(s - \frac{M}{2})$ by (ii), and (5.77) is at most $\big(MC_4 + C_2\big)\exp(-\epsilon_0)$ by (iv). Combining the above, we find that

$$\begin{aligned} V[\mathcal{W}(s),\mathcal{W}(t)] &\leq C_4(s - \frac{M}{2}) + (MC_4 + C_2)\exp(-\epsilon_0) \\ &= C_4 s + MC_4\Big((1 + \frac{C_2}{MC_4})\exp(-\epsilon_0) - \frac{1}{2}\Big). \qquad (5.82) \end{aligned}$$

Thus by (vi), $V[\mathcal{Z}(s),\mathcal{Z}(t)] - V[\mathcal{W}(s),\mathcal{W}(t)]$ is at least

$$\begin{aligned} &MC_4(\frac{1}{2} - (1 + \frac{C_2}{MC_4})\exp(-\epsilon_0) - \frac{3C_2}{MC_4}) \\ &= MC_4\Big(\frac{1}{2} - \exp(-\epsilon_0) - (\frac{C_2}{C_4}\exp(-\epsilon_0) + \frac{3C_2}{C_4})M^{-1}\Big) \\ &\geq MC_4(\frac{1}{2} - \exp(-\epsilon_0) - \frac{8(C_4+C_2+C_3)}{MC_4}) \geq 0\ , \qquad (5.83) \end{aligned}$$

by the definition of M. Combining (5.75), (5.81), and (5.83) completes the proof. □

Bibliography

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions.* 1972.

[2] R.J. Adler. An introduction to continuity, extrema, and related topics for Gaussian processes. *Inst. Math. Statist. Lecture Notes - Monograph Series*, 12, 1990.

[3] Z. Aksin, M. Armory, and V. Mehrotra. The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

[4] S. Asmussen. *Applied Probability and Queues, 2nd ed.* Springer, 2003.

[5] J.D. Atkinson and T.K. Caughley. Spectral density of piecewise linear first order systems excited by white noise. *Int. J. Non-Linear Mechanics*, (3):137–156, 1968.

[6] D. Berger, B. Daily, P.Dunn, D. Gamarnik, R. Levi, W.C. Levine, W.S. Sandberg, and T. Schoenmeyr. A model for understanding the impacts of demand and capacity on waiting time to enter a congested recovery room. *Anesthesiology*, 110(6):1293 - 304, 2009.

[7] P. Billingsley. *Convergence of Probability Measures.* 1999.

[8] A. Borodin and P. Salminen. *Handbook of Brownian Motion - Facts and Figures, 2nd edition.* Birkhauser Verlag, 2002.

[9] A.A. Borovkov. Some limit theorems in the theory of mass service, II. *Theor. Probability Appl.*, 10:375–400, 1965.

[10] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *Operations Research*, 52:17–34, 2004.

[11] B. Boukai. An explicit expression for the distribution of the supremum of Brownian motion with a change-point. *Purdue University Technical Report*, (88-40), 1988.

[12] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, S. Zeltyn, L. Zhao, and S. Haipeng. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.

[13] J. Browne, C. Heavy, and H.T. Papadopoulos. *Queueing Theory in Manufacturing Systems Analysis and Design.* Kluwer, New York, 1993.

[14] H. Buchholz. *The Confluent Hypergeometric Function.* Springer-Verlag, New York, 1969.

[15] T. Byczkowski, J. Malecki, and M. Ryznar. Hitting times of Bessel processes. *Preprint*, 2010.

[16] T. Byczkowski and M. Ryznar. Hitting distribution of geometric Brownian motion. *Studia. Math.*, 173(1):19–38, 2006.

[17] H. Callaert. Exponential ergodicity for birth-death processes. *Ph.D. Thesis, University of Louvain*, 1971.

[18] H. Callaert. On the rate of convergence in birth-and-death processes. *Bull.Soc.Math.Belg.*, 26:173–184, 1974.

[19] T. Cayilri and E. Veral. Outpatient scheduling in healthcare: A review of literature. *Production and Operations Management*, 12(4):519–548, 2003.

[20] C.S. Chang, J.A. Thomas, and S.H. Kiang. On the stability of open networks: a unified approach by stochastic dominance. *Queueing Systems*, 15:239–260, 1994.

[21] Y. Chao, C. Hsiung, and T. Lai. Extended renewal theory and moment convergence in Anscombe's theorem. *The Annals of Probability*, 7(2):304–318, 1979.

[22] M. Chen. Exponential l^2 convergence and l^2 spectral gap for markov processes. *Acta Mathematica Sinica, New Series*, 7(1):19–37, 1991.

[23] M. Chen. Estimate of exponential convergence rate in total variation by spectral gap. *Acta Mathematica Sinica*, 14(1):9–16, 1998.

[24] D. Chruscinski. Quantum mechanics of damped systems. II. damping and parabolic potential barrier. *Journal of Mathematical Physics*, 45(3):841–854, 2004.

[25] A. Clarke. A waiting line process of markov type. *The Annals of Mathematical Statistics*, 27(2):452–459, 1956.

[26] J.W. Cohen and O.J. Boxma. A survey of the evolution of queueing theory. *Statistica Neerlandica*, 39(2):143–158, 1985.

[27] D. Cox. *Renewal Theory.* Methuen and Co., 1970.

[28] J. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research.*, 35:347–362, 2010.

[29] D. J. Daley. Bounds for the variance of certain stationary point processes. *Stochastic Processes and their Applications*, 7(3):255 – 264, 1978.

[30] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer, 1998.

[31] P. Diaconis. The cutoff phenomenon in finite markov chains. *Proc. Nat. Acad. Sci.*, (4):1659–1664, 1996.

[32] D. Dominici. Asymptotic analysis of the Askey-scheme I: from Krawtchouk to Charlier. *Cent. Eur. J. Math.*, 5(2):280–304, 2007.

[33] D. Dominici. Asymptotic analysis of the Krawtchouk polynomials by the WKB method . *Ramananujan J.*, 15:303–338, 2008.

[34] J.L. Doob. The Brownian movement and stochastic equations. *The Annals of Mathematics*, 43(2):351–369, 1942.

[35] J.L. Doob. The elementary Gaussian processes. *Ann. Math. Statist.*, 15(3):229–282, 1944.

[36] N. Duffield and N. O'Connel. Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Camb. Phil. Soc.*, 118, 1995.

[37] T.M. Dunster. Uniform asymptotic expansions for Charlier polynomials. *J. Approx. Theory*, 112(1):93–133, 2001.

[38] A. Erdelyi, W. Magnus, F. Oberhettinger, F. Tricomi, and H. Bateman. *Higher Transcendental Functions, Volume II.* McGraw-Hill Book Company, New York, 1953.

[39] A.K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13, 1917.

[40] A.K. Erlang. *On the rational determination of the number of circuits.* The Copenhagen Telephone Company, Copenhagen, 1948.

[41] S. Ethier and T. Kurtz. *Markov processes: characterization and convergence.* Wiley and Sons, 2005.

[42] C. Fricker, P. Robert, and D. Tibi. On the rates of convergence of Erlang's model. *J. Appl. Probab.*, 36:1167–1184, 1999.

[43] D. Gamarnik and P. Momcilovic. Steady-state analysis of a multi-server queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 40:548–577, 2008.

[44] A. Ganesh, N. O'Connel, and D. Wischik. *Big Queues.* Springer-Verlag, Berlin, 2004.

[45] N. Gans, G. Koole, and A. Mandelbaum. Telephone call centers: tutorial, review, and research prospects. *Manufacturing and Service Operations Management*, 5(2):79–141, 2003.

[46] G. Gaukler, C. Li, R. Cannaday, S. Chirayath, and Y. Ding. Detecting nuclear materials smuggling: using radiography to improve container inspection policies. *Annals of Operations Research*, pages 1–23, 2010. 10.1007/s10479-010-0717-y.

[47] R.K. Getoor and M.J. Sharpe. Excursions of Brownian motion and Bessel processes. *Probability Theory and Related Fields*, 47:83–106, 1979.

[48] A. Going-jaeschke and M. Yor. A survey and some generalizations of Bessel processes. *Bernoulli*, 9:313–349, 2003.

[49] Y. Gordon. Some moment inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

[50] I.S. Gradshteyn and I.M.Ryzhik. *Table of Integrals, Series, and Products.* Academic Press, New York, 1980.

[51] L. Green, P. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

[52] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

[53] B.M. Hambly, G. Kersting, and A.E. Kyprianou. Law of the iterated logarithm for oscillating random walks conditioned to stay positive. *Stoch. Proc. Applic.*, 108:327–343, 2003.

[54] E. Hille. *Analytic Function Theory.* Ginn and Company, 1959.

[55] I. Ibragimov and Y. Rozanov. *Gaussian random processes.* 1978.

[56] D. Iglehart. Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability*, 2:429–441, 1965.

[57] D. Iglehart and W. Whitt. Multiple channel queues in heavy traffic I. *Advances in Applied Probability*, 2(1):150–177, 1970.

[58] R. Jackson. Some applications of queuing theory in operational research (1953 - 1989). *IMA Journal of Math Applied Business and Industry*, 2:127–140, 1989.

[59] D. Jagerman. Some properties of the Erlang loss function. *Bell System Techn. J.*, 53(3):525–551, 1974.

[60] P. Jelenkovic, A. Mandelbaum, and P. Momcilovic. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems: Theory and Applications*, 47(1-2):53–69, 2004.

[61] W. Kang and K. Ramanan. Asymptotic approximations for the stationary distributions of many-server queues. *Preprint.*, 2010.

[62] S. Karlin and J. McGregor. The differential equation of birth-and-death processes, and the Stieltjes moment problem. *Trans. Amer. math. Soc.*, 85:489–546, 1957.

[63] S. Karlin and J. McGregor. Many-server queueing processes with Poisson input and exponential service times. *Pacific J. Math.*, 8:87–118, 1958.

[64] H. Kaspi and K. Ramanan. SPDE limits of many server queues. *Preprint.*, 2010.

[65] J. Keilson. *Markov Chain Models: Rarity and Exponentiality.* Springer-Verlag, Berlin, 1979.

[66] O. Kella and W. Stadje. Superposition of renewal processes and an application to multi-server queues. *Statistics and Probability Letters*, 76(17):1914–1924, 2006.

[67] J. Kent. Some probabilistic properties of Bessel functions. *Annals of Probability*, 6(6):760–770, 1978.

[68] J. Kiefer and J. Wolfowitz. On the theory of queues with many servers. *Transactions of the American Mathematical Society*, 78(1):1–18, 1955.

[69] M. Kijima. Evaluation of the decay parameter for some specialized birth-death processes. *J.Appl.Prob.*, 29:781–791, 1992.

[70] C. Knessl and J. Leeuwaarden. Transient behavior of the Halfin-Whitt diffusion. *Preprint*, 2008.

[71] C. Knessl and J. Leeuwaarden. Spectral gap of the Erlang A model in the Halfin-Whitt regime. *Preprint*, 2010.

[72] W. Ledermann and G. Reuter. Spectral theory for the differential equations of simple birth and death processes. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences,*, 246(914):321–369, 1954.

[73] G. Lorden. On excess over the boundary. *The Annals of Mathematical Statistics*, 41(2):520 – 527, 1970.

[74] A. Mandelbaum and S. Zeltyn and. Call centers with impatient customers: Many-server asymptotics of the M/M/n + G queue. *Queueing Systems*, 51(3):361–402, 2005.

[75] A. Mandelbaum and P. Momcilovic. Queues with many servers: the virtual waiting-time process in the QED regime. *Math. Oper. Res.*, 33(3):561–586, 2008.

[76] A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189 – 1205, 2009.

[77] M. Marcus and J. Rosen. *Markov Processes, Gaussian Processes, and Local Times*. 2006.

[78] M.B. Marcus. Upper bounds for the asymptotic maxima of continuous Gaussian processes. *The Annals of Mathematical Statistics*, 43(2):522–533, 1972.

[79] V. Marianov and D. Serra. Location models for airline hubs behaving as M/D/c queues. *Computers and Operations Research*, 30(7):983 – 1003, 2003.

[80] J. Morrison and D. Martin. Practical extensions to cycle time approximations for the G/G/m queue with applications. *IEEE Transactions on Automation Science and Engineering*, 4(4), 2007.

[81] P. Morse. Stochastic properties of waiting lines. *Journal of the Operations Research Society of America*, 3(3):255–261, 1955.

[82] V. Piterbarg. *Asymptotic Methods in the theory of Gaussian processes and fields.* American Mathematical Society, 1996.

[83] J. Pitman and M. Yor. *Stochastic Integrals.* Springer, New York, 1981.

[84] J.W. Pitman. One-dimensional Brownian motion and the three-dimensional Bessel process. *Advances in Applied Probability*, 7(3):511–526, 1997.

[85] F. Pollaczek. U ber zwei formeln aus der theorie des wartens vor schaltergruppen. *Elektrische Nachrichtentechnik*, 8:256–268, 1931.

[86] N. Prabhu. *Stochastic Processes.* 1965.

[87] A. Puhalski and M. Reiman. The multiclass GI/PH/n queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(3):564–595, 2000.

[88] A.A. Puhalski and J. Reed. On many-server queues in heavy traffic. *Ann. Appl. Prob.*, 20(1):129–195, 2010.

[89] R.Bo and R. Wong. Uniform asymptotic expansion of Charlier polynomials. *Methods Appl. Anal.*, 1(3):294–313, 1994.

[90] J. Reed. The G/GI/N queue in the Halfin-Whitt regime II: idle time system equations. *Preprint.*, 2007.

[91] J. Reed. The G/GI/n queue in the Halfin-Whitt regime. *Annals of Applied Probability*, 19(6):2211–2269, 2009.

[92] J. Riordan. Moment recurrence relations for binomial, Poisson and hypergeometric frequency distributions. *Ann. Math. Statist.*, 8:103–111, 1937.

[93] S. Ross. *Stochastic Processes, 2nd ed.* Wiley and Sons, 1996.

[94] T. Saaty. Time-dependent solution of the many-server Poisson queue. *Operations Research*, 8(6):755–772, 1960.

[95] A. Selvi and M. Rosenshine. A queueing system for airport buses. *Transportation Research Part B: Methodological*, 17(6):427 – 434, 1983.

[96] V. Seshadri. Exponential models, Brownian motion, and independence. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 16(3):pp. 209–221, 1988.

[97] J. Shanthikumar and D. Yao. Stochastic monotonicity in general queueing networks. *J. App. Prob.*, 26:413–417, 1989.

[98] L.A. Shepp. The joint density of the maximum and its location for a Wiener process with drift. *Journal of Applied probability*, 16:423–427, 1979.

[99] D. Stoyan. *Comparison methods for queues and other stochastic models.* Wiley, 1983.

[100] W. Szczotka. Tightness of the stationary waiting time in heavy traffic. *Adv. Appl. Prob.*, 31:788–794, 1999.

[101] L. Takacs. *Introduction to the theory of queues.* Oxford University Press, New York, 1962.

[102] E.A. van Doorn. Stochastic monotonicity and queueing applications of birth death processes. *Lecture Notes in Statist.*, 4, 1981.

[103] E.A. van Doorn. On oscillation properties and the interval of orthogonality of orthogonal polynomials. *SIAM J. Math. Anal.*, 15:1031–1042, 1984.

[104] E.A. van Doorn. Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process. *Advances in Applied Probability*, 17:514–530, 1985.

[105] E.A. van Doorn. Representations and bounds for zeros of orthogonal polynomials and eigenvalues of sign-symmetric tri-diagonal matrices. *J.Approx.Theory*, 51:254–266, 1987.

[106] E.A. van Doorn and A.I. Zeifman. On the speed of convergence to stationarity of the Erlang loss system. *Queueing Systems*, 63:241–252, 2009.

[107] E.A. van Doorn, A.I. Zeifman, and T.L. Panfilova. Bounds and asymptotics for the rate of convergence of birth-death processes. *Teor. Veroyatnost. I Primenen.*, 54:18–38, 2009.

[108] L. Wein, A.H. Wilkins, M. Baveja, and S.E. Flynn. Preventing the importation of illicit nuclear materials in shipping containers. *Risk Analysis*, 26:1377–1393, 2006.

[109] W. Whitt. Comparing counting processes and queues. *Advances in Applied Probability*, 13(1):207–220, 1981.

[110] W. Whitt. Queues with superposition arrival processes in heavy traffic. *Stoch. Proc. App.*, 21:81–91, 1985.

[111] W. Whitt. Approximations for the GI/G/m queue. *Production and Operations Management*, 2(2):114–161, 1993.

[112] W. Whitt. The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution. *Queueing Syst. Theory Appl.*, 36(1/3):71–87, 2000.

[113] W. Whitt. *Stochastic Process Limits.* Springer, 2002.

[114] P. Whittle. Bounds for the moments of linear and quadratic forms in independent random variables. *Theor. Probability Appl.*, 5:302–305, 1960.

[115] D. Williams. Path decomposition and continuity of local time for one-dimensional diffusions, I. *Proc. London Math. Soc.*, 28:738–768, 1974.

[116] Shisheng Xie and Charles Knessl. On the transient behaviour of the Erlang loss model: heavy usage asymptotics. *SIAM J. Appl. Math.*, 53(2):555–599, 1993.

[117] Marc Yor. Some remarks about the joint law of Brownian motion and its supremum. In Jacques Azma, Marc Yor, and Michel Emery, editors, *Sminaire de Probabilits XXXI*, volume 1655 of *Lecture Notes in Mathematics*, pages 306–314. Springer Berlin / Heidelberg, 1997. 10.1007/BFb0119315.

[118] R. Young. Euler's constant. *The Mathematical Gazette*, (75):187–190, 1991.

[119] A.I. Zeifman. Some estimates of the rate of convergence for birth and death processes. *Journal of Applied Probability*, 28(2):268–277, 1991.