Essays on Health Care Operations Management

by

Amir Rastpour

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Operations and Information Systems

Faculty of Business

University of Alberta

ABSTRACT

This dissertation consists of three separate essays on health care operations management. Abstracts of the three essays are as follows:

Essay 1: We model emergency medical services (EMS) as Erlang loss systems and study ambulance shortage periods, intervals with few or no ambulances available to handle new emergency calls. We propose a simple recursion to calculate the expected duration of ambulance shortage periods and validate our recursion with data from Calgary, Canada, EMS. We develop analytical results for the second and higher moments, distribution, and Laplace transform of the shortage periods for some specific service time distributions. We provide analytical tools to investigate the impact of two possible actions that ambulance dispatchers can take: (1) requesting additional ambulances from neighbouring cities or other ambulance fleets, and (2) asking that busy ambulances be freed, for example the ones currently waiting to offload patients in EDs. Our models evaluate two performance measures: (1) the expected remaining duration of shortage periods, and (2) the expected number of lost calls.

Essay 2: Except for some special cases, closed form solutions for multi-server queues with multiple classes of impatient customers do not exist due to their high complexity. We model these systems as level-dependent quasi-birth-and-death (LDQBD) processes and propose two novel methods to numerically solve them: (1) we use Lyapunov analysis to truncate the state space such that the probability mass in the truncated upper tail is guaranteed to be smaller than a pre-specified value. This method can potentially substitute the currently-used heuristics that are exploited within algorithms that truncate the system first and then calculate its performance measures. (2) we extend an existing algorithm such that we can calculate the stationary probabilities with a desired error tolerance—current methods do not

provide bounds on the stationary probabilities.

Essay 3: We propose a tool to accurately predict the number of heart attack patients in sufficiently small geographical areas of Alberta. Focusing on small spatial units enables researchers to calculate precise estimates of travel times from the heart attack scene to a treatment center, which is useful in finding appropriate locations for new treatment facilities. We use standard multiple linear, Poisson, and negative binomial regression methods to predict the number of heart attacks as a function of the population in cohorts of age, sex, education, and income. We build, validate, and compare the performance of these methods using an empirical data set of heart attack counts in postal codes of Alberta from 2003 to 2010, and 2006 census data for Alberta dissemination areas.

In memory of my father, Kamal Rastpour.

ACKNOWLEDGEMENTS

I am sincerely grateful to my supervisor, Professor Armann Ingolfsson, for his continuous trust in my work, endless patience, and passion for teaching his deep knowledge to me. I also express my warmest gratitude to my other supervisor Dr. Padma Kaul, who was very supportive throughout my program.

I owe my deepest gratitude to Drs. Reidar Hagtvedt, Bora Kolfal, and Burhaneddin Sandıkçı for sharing their knowledge and experience with me to carry out my research. It would have been very difficult for me to finish my thesis if I did not have their continuous support. I also would like to take this opportunity to thank all members of the Operations and Information Systems department at the University of Alberta for their help throughout these years and my external examiners Drs. Edieal J. Pinker and Ming J. Zuo for their invaluable comments.

Finally, I am grateful for all the encouragement from my family and friends. This dissertation would not have been possible without them. Thank you especially to my mother, Faegheh Rezabakhsh, and my brother, Abbas Rastpour, for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF ALGORITHMS

Algorithm

CHAPTER 1

Introduction

This dissertation, written in partial fulfillment of the requirements for a Ph.D. degree in Operations Management in the Accounting, Operations, and Information Systems Department at the University of Alberta School of Business, consists of three separate papers on health care problems. These problems are associated with various aspects of the health care system, including emergency medical services (EMS), emergency departments, and demand prediction for a specific type of treatment.

The first paper, which is presented in Chapter 2 and has been co-authored by Drs. Armann Ingolfsson and Bora Kolfal, both from the Alberta School of Business, is on a problem in Ambulance fleet management. Mission-critical systems like fire, police, and EMS may experience disasters if they face capacity shortages. Therefore, it is necessary to have contingency plans to quickly restore these systems when their utilization goes up. We focus on EMS systems and study Red Alerts (when all ambulances are busy) and Yellow Alerts (when the number of available ambulances falls below a threshold). Possible actions that EMS dispatchers take during shortage periods include: (1) Requesting additional ambulances from neighbouring cities or other ambulance fleets, (2) asking to free up busy ambulances, for example the ones currently waiting to offload patients in EDs, or (3) repositioning available ambulances. In the EMS systems that we are familiar with, dispatchers decide on actions based on a combination of judgment and simple rules, such as the compliance tables for repositioning. The dynamics of EMS systems are sufficiently complicated

to make it difficult to reliably predict the consequences of different actions using unaided human judgment, even by highly experienced dispatchers. We focus on the first two actions and provide technical methods that EMS dispatchers can use to evaluate the impacts of the actions and make their decisions based on solid results of our methods.

The second paper, which is presented in Chapter 3 and has been co-authored by Dr. Armann Ingolfsson and Dr. Burhaneddin Sandıkçı from the University of Chicago's Booth School of Business, develops performance evaluation methods for two-class multi-server queues with abandonment. Such methods are useful, for example, for investigating ways to reduce emergency department (ED) waiting times. Prolonged waiting times in EDs have turned to serious problems that threaten patients health and lives throughout the world. As innovative solutions for long waiting times, hospitals redesign the patients' flow in EDs. Hospitals, for example, create a fast track for less acute patients through which these patients are seen by physician assistants without occupying resources needed by more acute patients. We provide an analytical tool to investigate the impact of such redesigns on waiting times of patients with different acuity levels. We view an ED as a queueing system with multiple servers (beds or physicians), multiple priority classes (acuity classes), and abandonment (patients who leave the system without being seen by a physician), and model it as an infinite level quasi-birth-and-death process. We specifically focus on queues with two impatient customer classes that have different service and patience rates. We use Lyapunov analysis and truncate the state space such that the stationary probability mass in the truncated upper tail of the state space is below some tolerance. As another analytical tool, we provide algorithms to calculate the steady state probabilities and performance measures with any desired accuracy. Our algorithm automatically truncates the state space such that the error tolerance is satisfied.

The third paper, which is presented in Chapter 4 and has been co-authored by Dr. Armann Ingolfsson, Dr. Reidar Hagtvedt from the Alberta School of Business, and Dr. Padma Kaul from the University of Alberta Faculty of Medicine &

Dentistry, is on a problem in demand prediction for heart attack treatment facilities. We propose a tool to accurately predict the number of heart attack patients in sufficiently small geographical areas of Alberta. Focusing on small spatial units enables researchers to calculate precise estimates of travel times from the heart attack scene to a treatment center, which is useful in finding appropriate locations for new treatment facilities. We use standard multiple linear, Poisson, and negative binomial regression methods to predict the number of heart attacks as a function of the population in cohorts of age, sex, education, and income. We build, validate, and compare the performance of these methods using an empirical data set of heart attack counts in postal codes of Alberta from 2003 to 2010, and 2006 census data for Alberta dissemination areas.

CHAPTER 2

Modeling Yellow and Red Alert Durations for Ambulance Systems

2.1 Introduction

Whether they are caused by a mass-casualty incident, extreme weather, a terrorist attack, heavy traffic, or an unanticipated demand surge, capacity shortages in mission-critical systems can lead to a disaster if no contingency plans have been made. These systems are designed to almost always have capacity to respond to emergencies. Still, periods when such systems run out of resources do happen and are important enough to have a name: Red Alerts. More generally, systems that provide highly time-sensitive services such as emergency services, intensive care units, and recovery services contractors for seminconductor manufacturing, aerospace, defense, medical equipment and other industries (Kim et al. 2010) are designed to minimize the frequency with which capacity is fully utilized, but reducing that frequency to zero could be impossible or prohibitively expensive. Surges in demand will occasionally cause all or nearly all capacity to be utilized. The purpose of this chapter is to propose and analyze models of the duration of such shortage periods and of actions that could be taken to reduce the duration or impact of such periods, focusing on the specific context of emergency medical services (EMS).

EMS practitioners distinguish between three levels of intensity of resource usage: High, medium, and low (Fitch et al. 1993). The highest intensity level corresponds to periods when no ambulances are available to cover medical emergencies and

these periods are typically referred to as "Red Alerts." A medium intensity level, corresponding to periods when less than a threshold number[1] of ambulances are available, is sometimes referred to as a "Yellow Alert," which is the terminology we will use in this chapter. Red Alerts pose health and safety issues that are of concern to the general public, as reflected in frequent media reports: Sinnema (2012) reports that the Edmonton, Alberta EMS system spent 9 hours and 45 minutes on Red Alert during the first nine months of 2010 (0.15% of the time), for example. Schneider (2012), Pedersen (2012), and myFOXdetroit (2011) are recent examples of reports in a similar vein, for the EMS systems in Calgary, Toronto, and Detroit, respectively.

Ambulance shortage periods commonly follow surges in emergency department (ED) workload that cause delays in freeing ambulances whose staff are waiting to transfer care of their patients to ED staff. Such delays tie up ambulances and reduce the EMS system's ability to respond to new calls. Sinnema (2010) reports that Edmonton paramedics were tied up in EDs for one hour and 22 minutes on average during the first 9 months of 2010. In order to shorten alert periods, EMS departments may ask EDs to prioritize the unloading of ambulances. The prioritization process has been formalized in an ED Surge Capacity Protocol in Alberta (Alberta Health Services 2010) and elsewhere (Stony Brook University Medical Center 2012, The College of Emergency Medicine 2012) and medical researchers have investigated the impact of such protocols on ED crowding (Cha et al. 2009, Watase et al. 2012).

In the EMS context, Yellow Alert periods are important from two perspectives: (1) the onset of a Yellow Alert is a signal to dispatchers to take actions to prevent the situation from deteriorating into a Red Alert, and (2) a small number of available ambulances in the system increases the average distance to the closest available ambulance, which results in longer response times. In practice, dispatchers take actions to attempt to shorten Yellow Alerts, but it is difficult to select actions because (1) it is difficult to predict Yellow Alert durations, and (2) there is a lack of appropriate performance measures that can be used to assess the likely impacts of an action. We will address both of these issues.

Table 2.1: EMS configuration in Edmonton,2008, and Calgary, 2009.

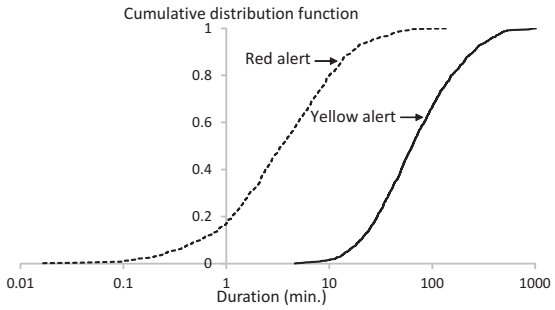
Parameter	Edmonton	Calgary
Yellow Alert threshold (θ)	8	12
Minimum number of scheduled ambulances	19	28
Maximum number of scheduled ambulances	36	54
Average number of scheduled ambulances	25	41
Average utilization	57%	43%

Table 2.2: Alert periods' descriptive stats. in Edmonton, 2008, and Calgary, 2009.

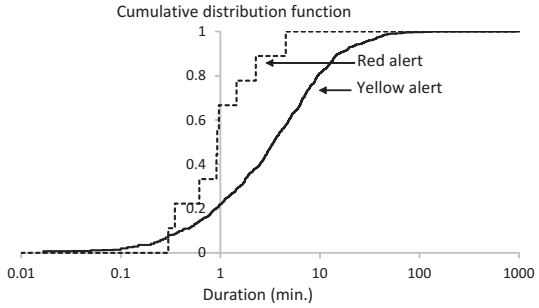
	Yellow Alert		Red Alert	
Statistic	Edmonton	Calgary	Edmonton	Calgary
Sample Size	1349	703	587	9
Mean (min.)	106.41	7.09	7.20	1.37
Standard Deviation (min.)	120.26	11.53	11.32	1.32
Maximum (min.)	1012.02	127.28	138.93	4.53
Squared Coefficient of Variation	1.28	2.64	2.47	0.94

Tables 2.1-2.2 and Figure 2.1 provide EMS configurations, descriptive statistics, and empirical distributions for EMS alert periods in Edmonton and Calgary during 2008 and 2009, respectively. Yellow and Red Alerts were more frequent in Edmonton than Calgary, consistent with the higher ambulance utilization in Edmonton. Alert period durations are highly variable (with squared coefficients of variation larger than one in most cases), which suggests that Red and Yellow Alert durations are difficult to predict.

Possible actions that EMS dispatchers take during shortage periods include: (1) Requesting additional ambulances, (2) asking to free up busy ambulances, for example the ones currently waiting to offload patients in EDs, or (3) repositioning available ambulances. Action (3) is only possible during Yellow Alerts but not during Red Alerts because there are no available ambulances to reposition. In the EMS systems that we are familiar with, dispatchers decide on actions based on a combination of judgment and simple rules, such as the compliance tables for repositioning that are discussed in Alanis et al. (2013). The dynamics of EMS systems are sufficiently complicated to make it difficult to reliably predict the consequences of different actions using unaided human judgment, even by highly experienced dispatchers. When dispatchers consider requesting new ambulances, for example,



(a) Edmonton



(b) Calgary

Figure 2.1: Alert periods' CDF in Edmonton, 2008, and Calgary, 2009.

as ambulance shortage durations are difficult to predict, they face the uncertainty of whether the shortage period will naturally end soon or whether it will last for an extended time period. Mobilizing new ambulances is costly and adds stress to dispatchers and ambulance crews, all to no avail, if the alert is short-lived.

One perspective is to view this problem as an optimal control problem and formulate it as a Markov decision process where costs of lost calls, requested ambulances, and expedited services are minimized. Instead of an MDP formulation, we use a simple $M/G/c/c$ model. Furthermore, we restrict the policy space to focus on policies similar to ones already used in practice, by assuming that the actions of requesting additional ambulances and increasing the service rate can only be taken if the system has reached a Yellow Alert period. We develop methods to quickly perform calculations to provide decision support to dispatchers, so that when the system is within a Yellow Alert, they can instantly see the impacts of triggering intervention actions on the system.

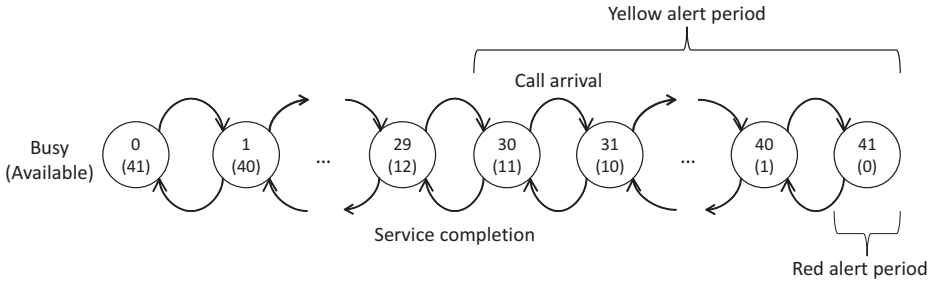


Figure 2.2: System states related to Yellow and Red Alerts. $c = 41$, $\theta = 12$.

We mathematically model the duration of alert periods and the impacts of requesting additional resources and freeing up busy ambulances on the severity and duration of these periods. The third possible action, repositioning, has been investigated recently by several researchers (Alanis et al. 2013, Maxwell et al. 2010, Schmid 2012). Our work complements this work by investigating other actions that EMS operators can take during ambulance shortage periods. Our models are intended to help compare the effectiveness of different actions.

We model EMS systems as Erlang loss ($M/G/c/c$) systems. We view calls that arrive during a Red Alert as "lost," as other researchers have done (Maxwell et al. 2010, for example), because typically other resources (the fire department or EMS supervisors, for example) respond to such calls, rather than the call waiting in a queue. We study Yellow and Red Alerts as special cases of "k-partial busy periods": time intervals during which k or more of the c servers are busy. Red Alerts are c-partial busy periods and Yellow Alerts are $(c - \theta + 1)$-partial busy periods, where θ is the Yellow Alert threshold for the number of busy ambulances—a Yellow Alert remains in effect with $\theta - 1$ or fewer available ambulances. Figure 2.2 illustrates Yellow and Red Alerts assuming $\theta = 12$ and $c = 41$.

We make the following contributions:

1. We prove an insensitivity result: The first moments of partial busy period durations depend on the service time distribution only through its mean and can be expressed in closed form. The higher moments are sensitive to the shape of the service time distribution.

2. We validate our model by showing that it can be used to predict the mean partial busy period durations for the Calgary EMS system.

3. We develop recursive equations for the Laplace Transform (LT) of partial busy period duration distributions for the $M/G/c/c$ system.

4. We characterize the distribution of partial busy period durations for loss systems with exponential ($M/M/c/c$) and generalized hyperexponential ($M/GH/c/c$) service time distributions. We analyze higher moments of partial busy period durations for $M/M/c/c$ systems and provide recursive solutions and monotonicity results for the variance and squared coefficients of variation of partial busy period durations.

5. We formulate and solve absorbing Markov chains to predict the impact of requesting new ambulances and freeing up ambulances in emergency departments on the residual duration of alert periods and on the number of lost calls during these residual periods.

6. We illustrate how our methods could be used to aid dispatchers in selecting a combination of actions that optimizes the residual duration of an alert period or the number of lost calls.

The remainder of the chapter is organized as follows. We review related past work in Section 2; we define and analyze k-partial busy period durations in Section 3; we validate our models in Section 4; we analyse the impacts of two actions, adding servers and increasing the service rate, on two performance measures, remaining alert period duration and the number of lost calls during these residual periods, in Section 5; we illustrate how ambulance dispatchers can use our methods to manage ambulance shortage periods in Section 6; and we conclude in Section 7.

2.2 Literature Review

We survey four streams of related literature: (1) modeling of EMS systems, (2) insensitivity results for loss systems, (3) modeling of partial busy periods, and (4)

strategies to mitigate capacity or inventory shortages in various contexts.

EMS System Models: We are not the first to model EMS systems as loss systems; see Restrepo et al. (2009) who model EMS systems as $M/G/c/c$ systems and Li and Whitt (2013) who treat them as more general loss systems, for example. The Erlang loss model ignores two key aspects of EMS systems, however: Servers (ambulances) are not homogeneous, because of their geographic locations, and parameters (arrival rates and number of servers) vary with time. Larson's (1974, 1975) exact and approximate hypercube queueing model (HQM) addresses the geographic heterogeneity of servers. Many researchers have used variants of HQM to study EMS systems. Researchers who have formulated queueing models to study ambulance repositioning (Alanis et al. 2013, Maxwell et al. 2010) and offload delay in EDs (Almehdawe et al. 2012) have also assumed that calls that arrive when all ambulances are busy are lost. Fewer researchers have explicitly incorporated time-varying parameters in an analytical EMS system model; Ignall and Walker (1977) did this for an EMS system and Kolesar et al. (1975) for police patrol cars. Simulation models of EMS systems typically do incorporate time-varying parameters (Henderson and Mason 2004, Mason 2013). We adopt the Erlang loss model for simplicity, in order to make progress on modeling the duration of partial busy periods and on modeling the impact of actions to mitigate capacity shortages. We assess the impact of some of the simplifications that are inherent in the Erlang loss model in Section 4.

Insensitivity results for loss systems: Taylor (2013) defines an insensitive stochastic model as one whose "stationary distribution depends on one or more of its constituent lifetime distributions only through the mean," and provides an extensive literature review. The best known insensitive stochastic models are $M/G/c/c$ and $M/G/\infty$.

Although the steady state probabilities of the $M/G/c/c$ system are insensitive to the service time distribution beyond its mean, the same is not true for the transient occupancy probabilities. We show that the first moments of the k-partial busy

period durations, although they are measures of transient behavior, are insensitive to the service time distribution beyond its mean.

Partial busy periods: Busy periods are unambiguously defined and well studied for single-server queues; they begin when a customer arrives to an empty system and last until the server becomes idle again for the first time. For analytical results, see Gross and Harris (1998, p. 102), for example. For multi-server queues, however, the terminology for busy periods varies. Omahen and Marathe (1978) use "busy period T_k" and Sharma (1990, Chap. 4.4) use "k-server busy period" to refer to k-partial busy periods. Artalejo and Lopez-Herrero (2001) use "partial busy periods" for what we refer to as 1-partial busy periods—that is, at least one server is busy—and they use "full busy period" for what we refer to as c-partial busy periods—that is, all servers are busy. Other authors (Chan et al. 2013, for example) have followed Artalejo and Lopez-Herrero in using the term "partial busy period," and we extend that term in defining k-partial busy periods.

Omahen and Marathe (1978) and Sharma (1990) studied k-partial busy periods for the $M/M/c$ and $M/M/c/N$ (with queue capacity $= N-c$) systems, respectively. Bountourelis et al. (2013) report that k-partial busy periods have not been studied for loss systems, except as a special case of the $M/M/c/N$ system. Our focus on loss systems allows us to obtain stronger results than those in Sharma (1990). Bountourelis et al. (2013) discuss applications of loss models in modeling hospital intensive care units (ICU) and highlight the importance of studying the length of periods during which ICUs are full, that is, c-partial busy period durations. We thoroughly investigate k-partial busy period durations, for $k = 1, \ldots, c$, for Erlang loss systems and provide formulas to calculate their LT, probability density function (PDF), and moments.

Shortage strategies: Alert periods are conceptually similar to low-inventory periods for a retailer or a manufacturer, or periods where almost all beds in a hospital ward are occupied. Lawson and Porteus (2000), Duran et al. (2004), and Veeraragha-

van and Scheller-Wolf (2008) discuss the use of "expediting" during low-inventory periods. Chan et al. (2011) discuss the use of "speedup" in an ICU in order to accommodate new patients that need to enter the ICU. Such short-term actions are not without risk—for example, KC and Terwiesch (2012) show that speedup can increase the chance of ICU readmission and decrease an ICU's peak capacity. The actions that we consider (requesting additional ambulances and freeing up busy ambulances) can be viewed as examples of expediting and speedup. We provide methods to compare the impacts of the two actions of interest on the expected residual Yellow Alert duration and the number of lost calls during this period.

2.3 Partial Busy Period Modeling

We model an EMS system as a multi-server loss system with Poisson arrivals and a general service time distribution, that is, as an $M/G/c/c$ queueing system. We use Q to denote a generic interarrival time (exponentially distributed with mean $1/\lambda$) and we use T to denote a generic service time (generally distributed with mean $1/\mu$) with cumulative distribution function (CDF) $F_T(t)$. In this section, we first present our results for an exponential service time distribution. We present generalizations of some of our results for state-dependent service rates and for deterministic (D), generalized hyperexponential (GH), and general continuous (G) service time distributions. See Table A.1 for a list of notations in this chapter.

A k-partial busy period is a period during which at least k out of c servers are busy. If an arrival at time t_0 increases the number of busy servers to k (decreases the number of available servers to $c-k$), then a k-partial busy period begins at t_0. This period ends when a departure leaves $k-1$ busy servers ($c-k+1$ available servers) behind for the first time after t_0. We use B_k to denote the length of a generic k-partial busy period and we use $\mathrm{E}(B_k)$, $\mathrm{Var}(B_k)$, $\mathrm{SCV}(B_k) = \mathrm{Var}(B_k)/\mathrm{E}(B_k)^2$, and $\mathscr{L}_{B_k}(s) = \int_0^\infty e^{-sx} f_{B_k}(x)\mathrm{d}x$ to denote the first moment, variance, squared coefficient of variation of B_k, and the LT of the PDF of B_k, $f_{B_k}(x)$, respectively.

Our mathematical approach relies on the Markov process state description:

$X(t) = \left(\nu(t), \tilde{T}_1(t), \ldots, \tilde{T}_{\nu(t)}(t)\right)$, where the right-continuous function $\nu(t) \in \{0, 1, \ldots, c\}$ is the number of busy servers at time t, and $\tilde{T}_1(t), \ldots, \tilde{T}_{\nu(t)}(t)$ is a random permutation of the residual service times of the busy servers at time t. Erlander (1967) proves that this process has the following stationary distribution:

$$\lim_{t\to+\infty} \Pr\left(\nu(t) = k, \tilde{T}_1(t) \le t_1, \ldots, \tilde{T}_k(t) \le t_k\right) = P_k \prod_{i=1}^{k} F_{\tilde{T}}(t_i), \tag{2.1}$$

where P_k is the probability of having k busy servers, and $F_{\tilde{T}}(t)$ is the stationary excess distribution of the service time, $F_{\tilde{T}}(t) = \mu \int_0^t (1 - F_T(s))\,\mathrm{d}s$. See Takács (1969) for a discussion of this result and Brumelle (1978) for a generalization. In what follows, we assume that the process starts with the stationary distribution (or equivalently, that the process started in the distant past).

All proofs are in Appendix A.

2.3.1 Partial Busy Periods for the $M/M/c/c$ System

In Theorem 2.1, we characterize the first two moments and the distributional form of k-partial busy period durations when the service time distribution is exponential.

Theorem 2.1. *If T is exponentially distributed, then B_k follows a hyperexponential (H) distribution with $c - k + 1$ components. The first moment, variance, squared coefficient of variation, and LT satisfy the following equations.*

$$\mathrm{E}(B_c) = \frac{1}{c\mu},\ \mathrm{E}(B_k) = \frac{\lambda}{k\mu}\mathrm{E}(B_{k+1}) + \frac{1}{k\mu},\ k = c-1, \ldots, 1. \tag{2.2}$$

$$\mathrm{Var}(B_c) = E(B_c)^2,\ \mathrm{Var}(B_k) = \frac{\lambda}{k\mu}\mathrm{Var}(B_{k+1}) + d_k,\ k = c-1, \ldots, 1, \tag{2.3}$$

$$d_k = \frac{\lambda}{k\mu}\mathrm{E}(B_{k+1})^2 + \mathrm{E}(B_k)^2.$$

$$\mathrm{SCV}(B_c) = 1,\ \mathrm{SCV}(B_k) = \frac{\lambda}{k\mu}\frac{\mathrm{E}(B_{k+1})^2}{\mathrm{E}(B_k)^2}(\mathrm{SCV}(B_{k+1}) + 1) + 1,\ k = c-1, \ldots, 1, \tag{2.4}$$

$$\mathscr{L}_{B_c}(s) = \frac{c\mu}{s + c\mu},\ \mathscr{L}_{B_k}(s) = \frac{k\mu}{\lambda + k\mu + s - \lambda\mathscr{L}_{B_{k+1}}(s)},\ k = c-1, \ldots, 1. \tag{2.5}$$

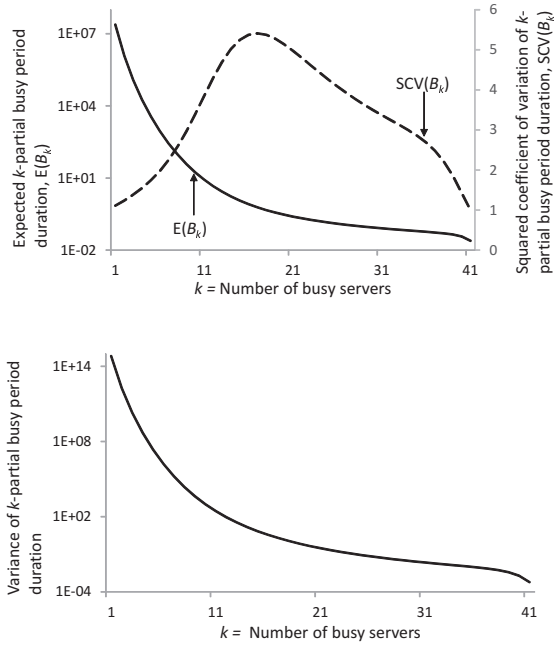


Figure 2.3: Numerical results for $M/M/41/41$ with $\lambda = 20$, $\mu = 1$.

The closed form solutions to the equations for $\mathrm{E}(B_k)$ *and* $\mathrm{Var}(B_k)$ *in* (2.2)-(2.3) *are as follows:*

$$\mathrm{E}(B_k) = \frac{1}{\mu} \sum_{i=0}^{c-k} \frac{(k-1)!}{(k+i)!} \left(\frac{\lambda}{\mu}\right)^i, \ k = 1, ..., c-1, \tag{2.6}$$

$$\mathrm{Var}(B_k) = \sum_{i=0}^{c-k-1} \frac{d_{k+i}(k-1)!}{(k-1+i)!} \left(\frac{\lambda}{\mu}\right)^i + \frac{(k-1)!}{(c\mu)^2(c-1)!} \left(\frac{\lambda}{\mu}\right)^{c-k}, \ k = 1, ..., c-1. \tag{2.7}$$

$\mathrm{E}(B_k)$ *and* $\mathrm{Var}(B_k)$ *are strictly decreasing in* k *but* $\mathrm{SCV}(B_k)$ *is not guaranteed to be monotonic in* k.

Figure 2.3 shows how $\mathrm{E}(B_k)$, $\mathrm{Var}(B_k)$, and $\mathrm{SCV}(B_k)$ vary with k for an $M/M/41/41$ system with $\lambda = 20$ and $\mu = 1$, and illustrates the monotonicity results in Theorem 2.1. We see that the SCV for B_k can be large (as high as 6 in this example), which is consistent with the high empirical SCV of 2.64 for yellow alert durations in Calgary that we saw in Section 1.

Equations (2.2)-(2.5) generalize to an $M/M/c/c$ system with state-dependent

service rates $\mu_k, k = 1, \ldots, c$, that is, the equations hold when we replace μ with μ_k. Erlang loss models with state-dependent service rates have applications in traffic flow modeling (Jain and Smith 1997), and in designing evacuation networks (Weiss et al. 2012). Alanis et al. (2013) indicate that ambulance service rates in an EMS system may depend on the number of busy ambulances.

2.3.2 Partial Busy Periods for the $M/GH/c/c$ System

The class of GH distributions is a generalization of the exponential distribution, to mixtures of exponential distributions where some of the mixture weights are allowed to be negative. The GH distribution is a useful modeling tool because the distribution of any positive random variable can be approximated with a GH distribution to any desired accuracy (Botta and Harris 1986). If the service time distribution is GH, then the expressions for the first moment in Theorem 2.1 remain valid, and we can characterize the class of distributions for the k-partial busy periods as generalized hyper-Erlang (GHE)—a mixture of Erlang distributions where some of the mixture weights are allowed to be negative.

Theorem 2.2. *Partial busy period durations of the $M/GH/c/c$ system follow GHE distributions, and their first moments can be calculated with* (2.2) *or* (2.6).

2.3.3 Partial Busy Periods for the $M/G/c/c$ System

The results from Theorem 2.1 regarding the first moment extend to general service time distributions, which implies that the first moments of k-partial busy periods in an $M/G/c/c$ system are insensitive to the shape of the service time distribution beyond its mean. This insensitivity property does not extend to higher moments of k-partial busy periods, however. Theorem 2.3 states these results formally, together with a recursive equation for the Laplace Transform of the PDF of B_k. Given k busy servers, we use the random variable R_k to denote the time between the *last* event (arrival or departure) epoch and the *next* event epoch, and we use events L_k (L_k^c) and N_k (N_k^c) to denote that the last event was an arrival (a

departure) and that the next event is an arrival (a departure), respectively.

Theorem 2.3. *The first moment,* $\mathrm{E}(B_k)$*, satisfies* (2.2) *and* (2.6)*, for general continuous and for deterministic service time distributions. The higher moments of* B_c*,* $\mathrm{E}(B_c^n)$ *for* $n \geq 2$*, are sensitive to the shape of the service time distribution. The LTs of the PDF for* B_k *satisfy the following recursion for a general continuous service time distribution:*

$$\mathscr{L}_{B_c}(s) = \mathscr{L}_{R_c|L_c}(s), \tag{2.8}$$

$$\mathscr{L}_{B_k}(s) = \mathscr{L}_{R_k|L_k}(s)\left(1 - \Pr(N_k|L_k)\frac{1 - \mathscr{L}_{R_k|L_k^c}(s)\mathscr{L}_{B_{k+1}}(s)}{1 - \Pr(N_k|L_k^c)\mathscr{L}_{R_k|L_k^c}(s)\mathscr{L}_{B_{k+1}}(s)}\right),$$

$$k = c-1, \ldots, 1, \tag{2.9}$$

Our derivations in Appendix A involve conditioning on whether a k-partial busy period will include one or more $(k+1)$-partial periods or none. We prove that the higher moments of B_c ($\mathrm{E}(B_c^n)$ for $n \geq 2$) are sensitive to the shape of the service time distribution, by demonstrating that $\mathrm{E}(B_c^n)$ is different for an exponential service time distribution than for a uniform service time distribution with the same mean. Extending the proof to $k < c$ does not appear to be easy but we conjecture that $\mathrm{E}(B_k^n)$ is sensitive to the service time distribution for $n \geq 2$, for $k = 1, \ldots, c-1$ just like it is for $k = c$. In Appendix A, we also prove (2.8)-(2.9) and provide expressions for their components.

2.4 Model Validation

In the preceding section, we analyzed the standard Erlang loss model, with extensions to state-dependent service rates. If we model an EMS system as a standard Erlang loss system, then we implicitly assume that the arrival rate, service rate, and number of servers are constants that do not vary with the time or the system state. In reality, arrival rates and the number of servers vary with time and past work (Alanis et al. 2013) suggests that service rates are state-dependent, which threatens the validity of the standard Erlang loss system. In this section, we investigate the

impacts of these deviations from our assumptions and propose a way to mitigate the negative impacts of these deviations.

We validate the Erlang loss model by estimating the model primitives (λ, μ, c) from data, using the primitives to compute *model outputs*, and comparing the model outputs to *empirical outputs*. The outputs that we focus on are the expected partial busy period durations, calculated using (2.2) (we use (2.2) with state-dependent service rates, μ_k, if the service rates are state-dependent). We proceed to compare the model outputs to empirical outputs in four steps: (1) We ignore the threats to the validity of the Erlang loss model and estimate non-time-varying and non-state-dependent model primitives. (2) We estimate state-dependent service rates μ_k. (3) We segment time into periods in which the assumptions of constant arrival rate and constant number of ambulances are more tenable. For each segment, we estimate non-time-varying and non-state-dependent model primitives. (4) Within each time segment obtained in (3), we estimate state-dependent service rates.

We use 2009 data from the Calgary, Alberta EMS systems. The data includes 108,420 calls. We removed 13,952 calls that were not followed by an ambulance dispatch and we removed an additional 734 calls because of missing or incorrect data, leaving 93,734 observations.

We find that time segmenting has a greater impact than using state-dependent service rates but both of these refinements to the analysis improve the agreement between model outputs and empirical outputs. We conclude this section by aggregating model outputs across time segments and comparing the aggregated predicted partial busy period durations to their empirical counterparts.

2.4.1 Step 1: Constant Parameters

We calculate the sample path for the number of busy ambulances $\nu(t)$ (right-continuous function $\nu(t) \in \{0, 1, \ldots, c\}$ is the number of busy servers at time t) by adding one at each call arrival epoch and subtracting one at each service completion epoch. We remove the data for 1 January 2009 and initialize the sample path with the number of active calls at 0:00 am on 2 January, based on the assumption

that none of these active calls arrived more than 24 hours before that instant. KC (2013) used a similar approach to initialize a sample path for the number of busy physicians in an emergency department. We use the sample path to compute samples of empirical k-partial busy period durations $\{b_{ki}, k = 1, \ldots, c, i = 1, \ldots, n_k\}$, which we use to calculate sample estimates $\overline{b}_k = (1/n_k) \sum_{i=1}^{n_k} b_{ki}$ of $\mathrm{E}(B_k)$, where b_{ki} and n_k are the ith k-partial busy period duration and the total number of k-partial busy periods in the sample path, respectively.

We estimated the arrival rate ($\hat{\lambda} = 10.69$ calls per hour) as the reciprocal of the average interarrival time, the service rate ($\hat{\mu} = 0.68$ patients per hour) as the reciprocal of the average service time, and the number of servers ($\hat{c} = 41$) as the rounded average number of scheduled ambulances in the data set. We then obtained the model outputs $\mathrm{E}(B_k)$ by using (2.2). Figure 2.4 shows large and systematic differences between the model outputs $\mathrm{E}(B_k)$ and the empirical outputs $\overline{b}_k$. In the remainder of this section, we reduce these differences by controlling for time of the week and for state-dependent service rates.

2.4.2 Step 2: State-dependent Service Rates

We follow Whitt (2012) in computing a death rate estimate $\hat{d}_k$ (and an associated 95% confidence interval) for state k as the number of transitions that reduce $\nu(t)$ from k to $k-1$, divided by the time spent in state k. The death rate for state k equals the number of busy servers, k, times the service rate per server, μ_k, and

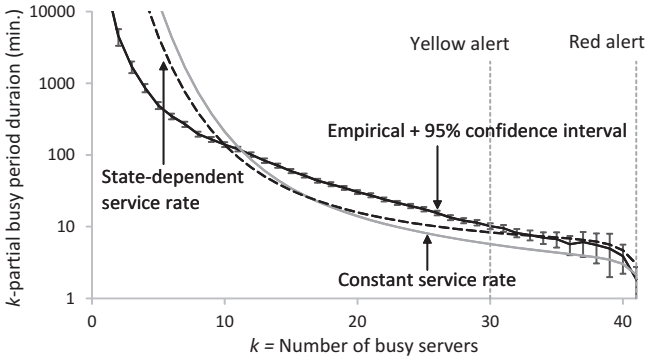


Figure 2.4: Predictions and empirical partial busy periods in Calgary, 2009.

therefore we estimate the state-dependent service rates as $\hat{\mu}_k = \hat{d}_k/k$.

As Figure A.3 shows, the estimated service rates decrease with the number of busy ambulances, which is consistent with findings in Alanis et al. (2013). They hypothesized that this "slowdown" effect occurs because a large number of ambulance patient arrivals causes ED crowding, which increases the time that ambulances are tied up in EDs, which translates to lower EMS service rates. We make a similar observation in our data that both average hospital times and average response times increase when the number of busy ambulances increase, which could be because of crowded EDs and increased travel distances; we further observe that the magnitude of increase in the hospital time is larger. We smoothed the service rate estimates using weighted linear regression (with the sample sizes for each k as weights), resulting in the linearly decreasing estimates $\hat{\mu}_k^{\text{linear}} = 0.86 - 0.0091k$ per hour. We combined these service rate estimates with our previous estimates for the arrival rate and the number of servers from Step 1 to obtain the state-dependent service rate model outputs in Figure 2.4. We obtained this curve by using (2.2) with μ_k. We see that incorporating state dependence improves the fit of the model outputs to the empirical outputs slightly, but large and systematic errors remain.

2.4.3 Step 3: Time-varying Parameters

EMS arrival rates are known to vary systematically by time of day and day of the week (Channouf et al. 2007, Setzler et al. 2009, Kim and Whitt 2014). Instead of explicitly incorporating time-varying parameters in our model (as Ignall and Walker (1977) did), we evaluate a simpler approach: We divide time into segments where the parameters do not vary much, and use our model separately for each time segment. This approach is similar to the "stationary-independent-period-by-period" (SIPP) approach discussed by Green et al. (2001).

Figure A.4 shows estimated hourly arrival rates with 95% confidence intervals and estimated hourly numbers of scheduled servers. Figures A.5-A.6 show estimated hourly arrival rates separately for "weekdays" (midnight on Sunday until 7 pm on Friday) and "weekends" (7 pm on Friday until midnight on Sunday). We chose 7

pm on Friday as a breakpoint because the arrival rate pattern is different for Friday, Saturday, and Sunday nights than for the rest of the week.

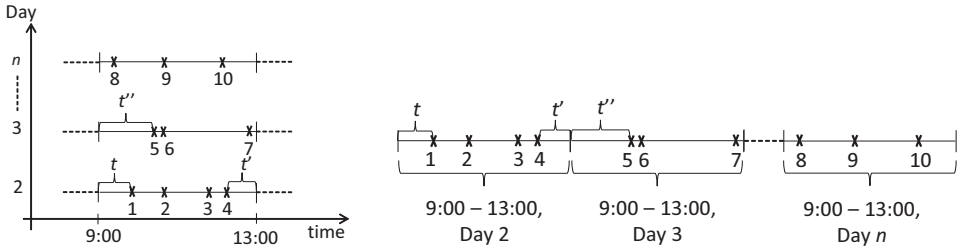
We divided every weekday and weekend day into 8 segments, namely 0:00-3:00, 3:00-7:00, 7:00-9:00, 9:00-13:00, 13:00-15:00, 15:00-19:00, 19:00-21:00, and 21:00-24:00. We describe the heuristic that we used to obtain the time segments in Appendix A.2, and we illustrate the segments in Figures A.5-A.6. Within each time segment, the arrival rate varies by at most 3 calls per hour and the number of servers varies by at most 5 ambulances. Although the arrival rates and the number of servers may not be close to constant within chosen time segments, as we shall discuss later in this section, their variation is small enough to meet our needs.

Many k-partial busy periods cross the boundaries between time segments, which complicates the task of obtaining empirical average partial busy period durations to compare to the model outputs. We address this by concatenating sample paths for a fixed time segment on consecutive days, as we illustrate in Figure 2.5. In order to explain this procedure, we focus on the weekday 9:00-13:00 segment (Segment 4) below:

1. Pool all calls with arrival epochs (but not necessarily service completion epochs) between 9:00 and 13:00 on weekdays in 2009, excluding 1 January (Figure 2.5(a)). Let n be the number of weekday 9:00-13:00 segments in the sample.

2. Calculate the inter-arrival time α_i for call i in the call pool (Figure 2.5(b)). For $i = 1$, the inter-arrival time is the time from 9:00 until its arrival ($\alpha_1 = t$). For the first call in the segment on day j, for $j = 3, \ldots, n$ the inter-arrival time is equal to the time from the last arrival epoch in the segment on day $j-1$ to 13:00 plus the time from 9:00 to the first arrival epoch in the segment on day j ($\alpha_5 = t' + t''$, for the example in Figure 2.5(b)).

3. Calculate the arrival epoch a_i and departure epoch d_i for each call i in the pool: $a_i = \sum_{n=1}^{i} \alpha_n$, $d_i = a_i + s_i$, where s_i is the service time of call i.



(a) Individual sample paths (b) Concatenated sample paths

Figure 2.5: We concatenate sample paths for 9:00 - 13:00 weekday calls.

4. Starting from the number of active calls at 9:00 on 2 January 2009, construct a sample path $\nu(t)$ for the number of busy ambulances by incrementing at every arrival epoch and decrementing at every service completion epoch.

5. Compute k-partial busy period durations and their sample averages, as previously described.

We estimated constant model primitives $\hat{\lambda}^{(\tau)}$, $\hat{\mu}^{(\tau)}$, and $\hat{c}^{(\tau)}$, separately for each time segment $\tau = 1, \ldots, 16$, and used these primitives to compute model outputs for each segment by using the method in Step 1. We obtained $\hat{\lambda}^{(4)} = 13.37$ calls per hour, $\hat{\mu}^{(4)} = 0.58$ patients per hour, and $\hat{c}^{(4)} = 42$ ambulances for weekdays 9:00-13:00 (time segment $\tau = 4$), for example. As Figure 2.6 illustrates, we obtain an excellent fit between model and empirical outputs for Segment 4. Graphs for the other 15 segments show excellent fits as well.

Kim and Whitt (2014) propose more sophisticated methods than we used to obtain segments within which the arrival rate is almost constant, but as we saw, the segments we chose are sufficient to obtain close agreement between model and empirical outputs, even though the arrival rates (and the number of servers) are not close to constant within each of our segments.

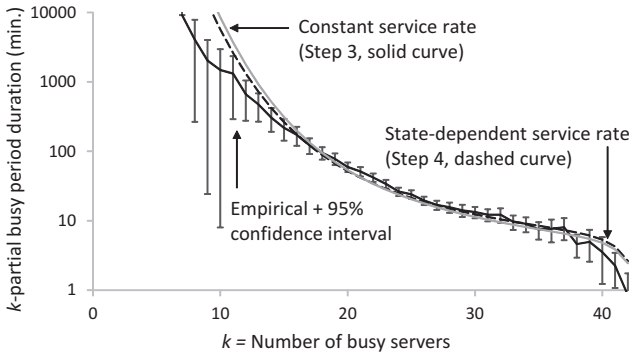


Figure 2.6: Model outputs for weekday 9:00 to 13:00.

2.4.4 Step 4: State- and Time-dependent Parameters

As a final step, we estimate state-dependent service rates $\mu_k^{(\tau)}$ separately for each time segment τ and state k and use them, together with segment-specific estimates for arrival rate and number of servers, obtained in Step 3, to compute the model outputs. As Figure 2.6 illustrates, incorporating state-dependent service rates has only a minor impact over Step 3 for the weekday 9:00-13:00 time segment, and the same was true for the other time segments as well.

2.4.5 Aggregating Over Time Segments

Using the Step 4 analysis, we calculate expected partial busy period durations $\mathrm{E}\left(B_k^{(\tau)}\right)$ for every time segment $\tau = 1, \ldots, 16$ and $k = 1, \ldots, \hat{c}^{(\tau)}$. We aggregate over the time segments in order to obtain model outputs that, we hope, will closely match the empirical outputs $\bar{b}_k$ for the whole sample.

Let B'_k be the duration of a k-partial busy period in the real EMS system, which as we have seen has time-varying arrival rates, server counts, and state-dependent service rates. We proceed heuristically to derive an expression for $\mathrm{E}(B'_k)$ as a function of the $\mathrm{E}\left(B_k^{(\tau)}\right)$ that we obtained from homogeneous Erlang loss models

with state-dependent service rates:

$$\mathrm{E}\left(B'_k\right) \approx \sum_{\tau=1}^{16} \Pr(B'_k \text{ begins in time segment } \tau)\mathrm{E}\left(B_k^{(\tau)}\right) \approx \frac{\sum_{\tau=1}^{16} \mathrm{E}\left(N_k^{(\tau)}\right) \mathrm{E}\left(B_k^{(\tau)}\right)}{\sum_{\tau=1}^{16} \mathrm{E}\left(N_k^{(\tau)}\right)}, \tag{2.10}$$

where $N_k^{(\tau)}$ is the number of k-partial busy periods that begin in the segment-τ sample path. We approximate $\mathrm{E}\left(N_k^{(\tau)}\right)$ as:

$$\mathrm{E}\left(N_k^{(\tau)}\right) = l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}, \tag{2.11}$$

where $l^{(\tau)}$ is the total time spent in Segment τ, and $\pi_{k-1}^{(\tau)}$ is the steady-state probability that the segment-τ Erlang loss system has $k-1$ busy servers. We use standard formulas (Gross and Harris 1998, p. 80, for example) to obtain $\pi_{k-1}^{(\tau)}$. The rationale for the approximation is that $\lambda^{(\tau)}\pi_{k-1}^{(\tau)}$ is the steady-state rate at which k-partial busy periods begin in the segment-τ system. Combining approximations (2.10)-(2.11), we obtain the following weighted average model outputs:

$$\mathrm{E}(B'_k) = \sum_{\tau=1}^{16} w_k^{(\tau)}\mathrm{E}\left(B_k^{(\tau)}\right), \; w_k^{(\tau)} = \frac{l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}}{\sum_{\tau=1}^{16} l^{(\tau)}\lambda^{(\tau)}\pi_{k-1}^{(\tau)}}.$$

As Figure 2.7 shows, these weighted average model outputs provide a good fit to the empirical outputs $\bar{b}_k$, as 2/3 of $\mathrm{E}(B'_k)$s are within the confidence intervals for k values of 30 or higher, which correspond to alert periods. It is noteworthy that this excellent fit is obtained even though we ignore the spatial distribution of ambulances, calls, and hospitals in our model.

2.5 Modeling the Impact of Operational Changes

As discussed in Section 3.1, once a Yellow Alert period begins, ambulance dispatchers face the uncertainty of whether the shortage period will end soon naturally, or whether the system will operate with a shortage of available ambulances for an extended period of time that could lead to a Red Alert. In this section, we extend

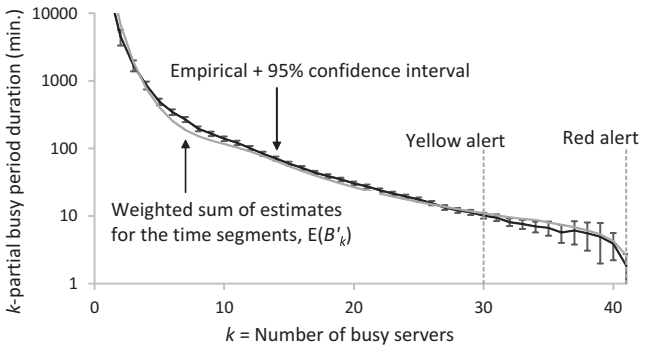


Figure 2.7: Aggregating over time segments.

the Erlang loss model to incorporate two corrective actions that dispatchers could take to reduce the duration of a Yellow Alert and the number of lost calls during Red Alerts: (1) Call in additional servers (ambulances) and (2) increase the service rate.

Call in additional servers: In a real system, it might be possible to request ambulances from neighboring municipalities, from another service (from an interfacility-transfer ambulance fleet, for example), or by asking new ambulance crews to come on duty. We model this action by adding a parameter n for the number of additional ambulances and a parameter $1/\delta$ for the expected value of the exponentially distributed time V until all n additional ambulances are available; we assume that all n new ambulances become available at the same time.

Increase the service rate: During extreme ambulance shortages, service times can be reduced by *expediting*, which could take the form of prioritizing the unloading of ambulances that are tied up in hospital EDs. Such prioritization should shorten the service times of those busy ambulances that are currently in EDs or will soon arrive at an ED and, therefore, the rate at which busy ambulances become available should increase. We model the service rate increase to capture, at least roughly, the overall effects of expediting. We add a single parameter to model this action—the new and increased service rate, μ_{new}; assumed to take effect as soon as the action is taken.

In reality, these actions would presumably be reversed after a Yellow Alert has

ended and the two actions we consider may have different flexibilities in terms of getting reversed as crew members that have been called in may need to stay on work until the end of their shift, but it would be easier for ED personnel to reverse their actions and offload ambulance patients with normal speed. We do not specify when the actions are reversed, in order to keep our models simple and because our main interest is in how these actions impact the duration and severity of a Yellow Alert.

We need performance measures to quantify the impact of the two actions. We know of no formal performance measures that are used in practice. The outcome that dispatchers would like to avoid, however, is a call that arrives during a Red Alert. We define two performance measures that are related to this outcome: (1) the expected remaining Yellow Alert duration, and (2) the expected number of lost calls during the remaining Yellow Alert duration. The first measure is an easy-to-interpret proxy for the outcome of interest. The second measure is directly related to the outcome of interest. If we are within a k-partial busy period and the current number of busy servers is $k' \geq k$, then we use $\tilde{B}_{kk'}$ to denote the remaining k-partial busy period duration and $H_{kk'}$ to denote the number of lost calls during the remaining k-partial busy period duration. Setting k equal to $c - \theta + 1$, where θ is the Yellow Alert threshold, provides our two performance measures as a function of the current number of busy servers, as $\mathrm{E}\left(\tilde{B}_{kk'}\right)$ and $\mathrm{E}\left(H_{kk'}\right)$.

We begin by computing $\mathrm{E}\left(\tilde{B}_{kk'}\right)$ and $\mathrm{E}\left(H_{kk'}\right)$ assuming that no action is taken, for an $M/M/c/c$ system. Then we extend the analysis to include the impacts of the two actions. We use standard results from the theory of absorbing continuous-time Markov chains. We indicate how the results generalize to an $M/G/c/c$ system by analysing the imbedded Markov chain at event epochs (arrival, departure, and the availability of new servers if we request new servers), using standard results for absorbing discrete-time Markov chains. In this section, we use Ω to denote the state space for an absorbing Markov chain, A to denote the set of absorbing states, and A^c to denote the set of transient states (with $A \cup A^c = \Omega$).

2.5.1 The $M/M/c/c$ System

For the $M/M/c/c$ system, we calculate the performance measures at an arbitrary time epoch t_0 within a k-partial busy period, given no action; given that we request additional servers; and given that we increase the service rate. We assume that the number of busy servers at t_0 is $\nu(t_0) = k'$, $k' = k, \ldots, c$. For convenience, we suppress the dependence on time.

2.5.1.1 No Action

We decompose the residual duration as $\tilde{B}_{kk'} = \Upsilon_{k'} + \Upsilon_{k'-1} + \cdots + \Upsilon_k$, where Υ_i is the time it takes for the number of busy servers to decrease from i to $i-1$. For the $M/M/c/c$ system, because of the memoryless property of the exponential distribution, Υ_i equals B_i in distribution. Therefore:

$$\mathrm{E}\left(\tilde{B}_{kk'}\right) = \sum_{i=k}^{k'} \mathrm{E}(B_i),\ k' = k, \ldots, c, \tag{2.12}$$

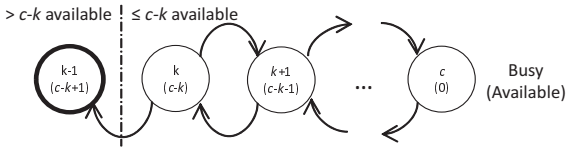
To compute the expected number of lost calls before the current Yellow Alert ends, we modify the $M/M/c/c$ system such that $\Omega = \{k-1, ..., c\}$ and $A = \{k-1\}$, as depicted in Figure 2.8(a). The infinitesimal generator matrix Q in canonical form (Kao 1996) is

$$Q = \begin{array}{c|cc} & A & A^c \\ \hline A & 0 & 0 \\ A^c & Y & Z \end{array}. \tag{2.13}$$

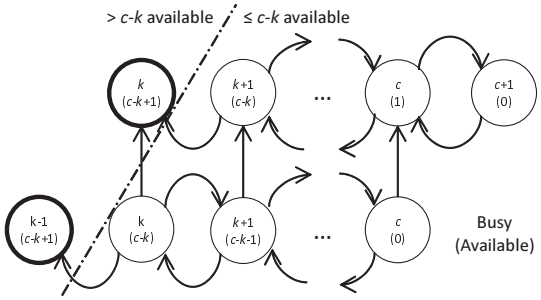
The fundamental matrix (Kao 1996, p. 256) for this Markov chain is

$$V = -Z^{-1}, \tag{2.14}$$

where v_{ij} is the expected time spent before absorption in transient state j, given that the chain begins in transient state i. The fundamental matrix provides an alternative way to obtain $\mathrm{E}\left(\tilde{B}_{kk'}\right)$, as $\sum_{i \in A^c} v_{k'i}$ (the total expected time spent in



(a) No action is taken.



(b) A new server is added.

Figure 2.8: Absorbing states (indicated by thicker borders) in modified systems.

all transient states prior to absorption). The fundamental matrix also provides a way to obtain the expected number of lost calls, which equals the call arrival rate times the expected time spent in state c, that is:

$$\mathrm{E}(H_{kk'}) = \lambda v_{k'c}, \quad k = 1, \ldots, c, \; k' = k, \ldots, c. \tag{2.15}$$

2.5.1.2 Add Servers

We augment the $M/M/c/c/$ state space by adding the indicator state variable $w(t)$ for whether the n requested servers have become available and use the ordered pair $(\nu(t), w(t))$ to denote the state at time t. We also define an *adjusted* k-partial busy period, which begins when the number of busy servers increases to k and ends when the number of *available* servers increases to more than $c-k$ for the first time.

Figure 2.9 illustrates the difference between adjusted and regular k-partial busy periods when $n = 1$. Both of these periods begin when there are c scheduled servers in the system and the number of busy servers increases to k. The regular partial busy period ends when the system enters a state with less than k busy servers (left of the dashed line) while the adjusted one ends when the system enters a state with

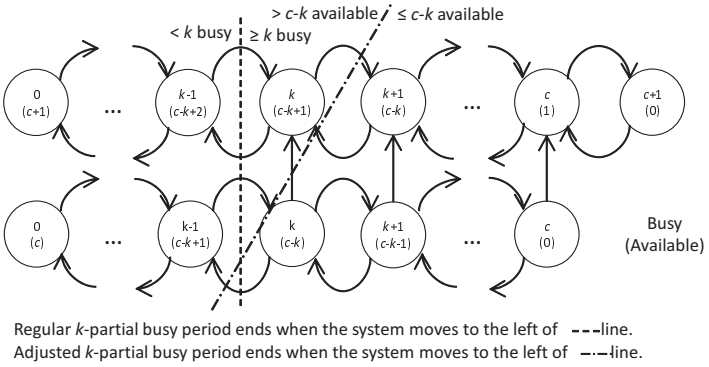


Figure 2.9: Regular and adjusted k-partial busy periods when $n = 1$.

more than $c - k$ available servers (left of the dashed-dot line).

We modify the Markov chain $(\nu(t), w(t))$ such that $\Omega = \{(k-1,0), ..., (c,0), (k,1), ..., (c+n,1)\}$ and $A = \{(k-1,0), (k,1), ..., (k+n-1,1)\}$. Figure 2.8(b) shows Ω, A, and A^c when $n = 1$. The expected residual duration of the adjusted k-partial busy period equals the total time spent before absorption in all transient states:

$$\mathrm{E}\left(\tilde{B}_{kk'}\right) = \sum_{s \in A^c} v_{(k',0)s}, \quad k = 1, \ldots, c, \ k' = k, \ldots, c, \tag{2.16}$$

We obtain the expected number of lost calls using the same logic as for (2.15), except now we have two Red-Alert states—$(c,0)$ and $(c+n,1)$:

$$\mathrm{E}\left(H_{kk'}\right) = \lambda\left(v_{(k',0)(c,0)} + v_{(k',0)(c+n,1)}\right), \ k = 1, \ldots, c, \ k' = k, \ldots, c. \tag{2.17}$$

2.5.1.3 Increase the Service Rate

When we increase the service rate from μ to $\mu_{\text{new}} = a\mu$ $(a > 1)$, we assume that the remaining service times for all calls currently in service, as well as the service times for all new calls, will be exponentially distributed with rate μ_{new}. We therefore re-evaluate (2.12) and (2.15) using μ_{new} to evaluate the impact of a service rate change on the expected residual k-partial busy period duration and on the expected number of lost calls, respectively.

2.5.1.4 Take Both Actions at the Same Time

To analyze the impacts of calling n new ambulances in and increasing the service rate form μ to μ_{new}, we re-evaluate (2.16)-(2.17) using μ_{new}.

2.5.2 The $M/G/c/c$ System

In analyzing the $M/M/c/c$ system, we allowed the time when a corrective action is taken to be an arbitrary epoch within a k-partial busy period. In the $M/G/c/c$ system, however, we assume that the corrective action is taken at an arrival epoch, t_0, within a k-partial busy period, with $\nu(t_0) = k' \in \{k+1, \ldots, c\}$ busy servers immediately after the arrival.

Our analysis depends on t_0 only through k' and therefore we normally suppress dependence on t_0 in our notation. A similar analysis can be done for a corrective action at a departure epoch. We omit the details, both for brevity and because we expect corrective actions to be less relevant at a departure epoch, when an ambulance becomes available without any dispatching action.

2.5.2.1 No Action

Similar to Section 2.5.1.1, we decompose the residual duration as $\tilde{B}_{kk'} = \Upsilon_{k'} + \Upsilon_{k'-1} + \cdots + \Upsilon_k$. As we assume that t_0 is an arrival epoch, $\Upsilon_{k'}$ is equal in distribution to $B_{k'}$. The next component, $\Upsilon_{k'-1}$, is the residual duration of a $(k'-1)$-partial busy period, starting immediately after a departure that leaves $k'-1$ busy servers behind, that is, $\Upsilon_{k'-1}$ is distributed as $\tilde{B}_{k'-1k'-1}|L^c_{k'-1}$. More generally, Υ_i is distributed as $\tilde{B}_{ii}|L^c_i$ for $i = k'-1, k'-2, ..., k$. Therefore,

$$\mathrm{E}\left(\tilde{B}_{kk'}|L_{k'}\right) = \mathrm{E}\left(B_{k'}\right) + \sum_{i=k}^{k'-1} \mathrm{E}\left(\tilde{B}_{ii}|L^c_i\right), \quad k = 1, \ldots, c,\ k' = k+1, \ldots, c, \tag{2.18}$$

We use Theorem 2.4 to compute terms in (2.18).

Theorem 2.4. *In the $M/G/c/c$ system, the following equations hold:*

$$\mathrm{E}\left(\tilde{B}_{ii}|L_i^c\right) = \frac{\mathrm{E}(R_i|L_i^c)}{1-\lambda\mathrm{E}(R_i|L_i^c)}(1+\lambda\mathrm{E}(B_{i+1})), \quad i = c-1, c-2, \ldots, 1, \tag{2.19}$$

where $\mathrm{E}(R_i|L_i^c)$, the expected sojourn time when there are k busy servers in the system and the last event was a departure, is calculated by:

$$\mathrm{E}(R_i|L_i^c) = \int_0^\infty e^{-\lambda t}(1-F_{\tilde{T}}(t))^i \,\mathrm{d}t.$$

To compute the expected number of lost calls before the current Yellow Alert ends, we recall the state description used in Section 2.3: $X(t) = \left(\nu(t), \tilde{T}_1(t), \ldots, \tilde{T}_{\nu(t)}(t)\right)$. Erlander (1967) shows that the stationary distribution for $X(t)$ is of product form (see (2.1)), meaning that as t tends to infinity, the residual service times of the busy servers, $\tilde{T}_i$, are i.i.d. and follow the stationary excess distribution, $F_{\tilde{T}}(t)$. Immediately after a departure, the residual service times remain i.i.d. Immediately after an arrival, the residual service times also remain i.i.d., except that the remaining service time of the newly arrived customer follows the service time distribution, $F_T(t)$. We use these results to define the following Markov chain that is imbedded immediately after arrival and departure epochs: (ν_n, e_n), where $\nu_n = \nu(t_n^+)$ and e_n is a binary variable equal to 1 if the n-th event at epoch t_n is an arrival and 0 if it is a departure.

We modify the imbedded discrete-time Markov chain of the $M/G/c/c$ system such that $\Omega = \{(k-1,0), ..., (c-1,0), (k+1,1), ..., (c,1)\}$ and $A = \{(k-1,0)\}$. The transition probability matrix P in canonical form (Kao 1996) is

$$P = \begin{array}{c|cc} & A & A^c \\ \hline A & I & 0 \\ A^c & Y & Z \end{array}. \tag{2.20}$$

The fundamental matrix (Kao 1996, p. 188) for this Markov chain is

$$W = (I - Z)^{-1}, \tag{2.21}$$

where w_{ij} is the expected number of times the system visits the transient state j before absorption, given that the chain begins in transient state i. The fundamental matrix provides an alternative way to obtain the expected residual Yellow Alert duration, by multiplying the expected number of visits to each transient state with the expected time spent per visit to that state, summed over all transient states:

$$\mathrm{E}\left(\tilde{B}_{kk'}|L_{k'}\right) = \sum_{s\in A^c} \left(w_{(k',1)s}\right) \mathrm{E}(R_s), \quad k = 1,\ldots,c,\ k' = k+1,\ldots,c. \tag{2.22}$$

We calculate expected sojourn times $\mathrm{E}\left(R_{(c,1)}\right)$, $\mathrm{E}\left(R_{(.,1)}\right)$, and $\mathrm{E}\left(R_{(.,0)}\right)$ as in (A.4), (A.13) and (A.18), respectively. The fundamental matrix also provides a way to obtain the expected number of lost calls, which equals the call arrival rate times the expected number of times the chain visits state c times the expected time spent in state c per visit:

$$\mathrm{E}\left(H_{kk'}|L_{k'}\right) = \frac{\lambda w_{(k',1)(c,1)}}{c\mu}, \quad k = 1,\ldots,c,\ k' = k+1,\ldots,c. \tag{2.23}$$

2.5.2.2 Add Servers

As in 2.5.1.2, we augment the state space of the imbedded discrete-time Markov chain (ν_n, e_n) to (ν_n, e_n, g_n), by adding an indicator state variable g_n for whether the n requested servers have become available, and we modify the imbedded Markov chain such that $\Omega = \{(k-1,0,0),...,(c-1,0,0),(k+1,1,0),...,(c,1,0),(k,0,1),...,(c+n-1,0,1),(k+1,1,1),...,(c+n,1,1)\}$, and $A = \{(k-1,0,0),(k,0,1),...,(k+n-1,0,1)\}$. Figure 2.8(b) shows Ω, A, and A^c when $n = 1$. We extend (2.16) to obtain $\mathrm{E}\left(\tilde{B}_{kk'}\right)$:

$$\mathrm{E}\left(\tilde{B}_{kk'}|L_{k'}\right) = \sum_{s\in A^c} w_{(k',1,0)s}\mathrm{E}(R_s), \quad k = 1,\ldots,c,\ k' = k+1,\ldots,c. \tag{2.24}$$

We show how to calculate expected sojourn times, E(R_s)s, in Appendix A.5.

We extend (2.17) to obtain E $(H_{kk'})$:

$$\mathrm{E}\left(H_{kk'}|L_{k'}\right) = \lambda\left(\frac{w_{(k',1,0)(c,1,0)}}{c\mu} + \frac{w_{(k',1,0)(c+n,1,1)}}{(c+n)\mu}\right),$$
$$k = 1,\ldots,c,\ k' = k+1,\ldots,c. \tag{2.25}$$

2.5.2.3 Increase the Service Rate

Similar to 2.5.1.3, we re-evaluate (2.18) and (2.23) using μ_{new} to see the impacts of the service rate change on the expected residual k-partial busy period durations and the expected number of lost calls, respectively.

2.5.2.4 Take Both Actions at the Same Time

To analyze the impacts of calling n new ambulances in and increasing the service rate form μ to μ_{new}, we re-evaluate (2.24)-(2.25) using μ_{new}.

2.6 Numerical Results and Managerial Insights

We illustrate our results using a scenario that approximates the Calgary EMS system on weekdays from 9:00 to 13:00 (Segment 4). We investigate the impacts of several parameters on the effectiveness of possible actions. We formulate two optimization problems, which we solve via complete enumeration, in order to display the tradeoffs between the two actions that we consider. We implemented the methods discussed in Section 2.5 in Matlab. The calculations were, in all cases, near-instantaneous, making the methods suitable for inclusion in real-time decision support systems.

We begin by relating the abstract action of increasing the service rate from μ to $\mu_{\text{new}} = a\mu$ to the concrete action of freeing of ambulances tied up in EDs. We decompose the number of busy ambulances as $k' = k_1 + k_2 + k_3$ where k_1 is the number of ambulances in EDs that will be released early, within the next $1/\mu_{\text{early}}$ ($\mu_{\text{early}} > \mu$) time units; k_2 is the number of ambulances in EDs that will be released

within the next $1/\mu$ time units; and k_3 is the number of ambulances that are busy outside of EDs and work with the service rate μ. We approximate the average service time after taking the action of freeing k_1 ED ambulances within $1/\mu_{\text{early}}$ time units as

$$\frac{1}{\mu_{\text{new}}} = \frac{1}{a\mu} = \frac{k_1}{k'}\frac{1}{\mu_{\text{early}}} + \frac{k_2 + k_3}{k'}\frac{1}{\mu}, \tag{2.26}$$

which implies that

$$a = \frac{k'}{(\mu/\mu_{\text{early}})k_1 + k_2 + k_3}. \tag{2.27}$$

It is inherent in this approximation that we assume that the remaining service times of all k' busy ambulances and the service times of new calls are exponentially distributed with rate μ_{new}.

The parameter estimates for Segment 4 are: $\hat{\lambda}^{(4)} = 13.37$ calls per hour, $\hat{\mu}^{(4)} = 0.58$ patients per hour, and $\hat{c}^{(4)} = 42$ ambulances. We focus on a situation where the number of busy servers is $k' = 40$ (2 ambulances available), and we study the sensitivity of the performance measures to different actions that dispatchers take. We also find the best combination of corrective actions to take under given conditions.

2.6.1 Sensitivity Analysis

Figure 2.10 illustrates the sensitivity of the two performance measures to the number of ambulances freed from EDs ($k_1 = 1, 2, 3$) and the time it takes to free them ($1/\mu_{\text{early}} = 0, 5, ..., 100$ minutes). For a given number of freed ambulances, we observe that both performance measures improve linearly with the time it takes to free the ambulances. The marginal impact of increasing the number of ambulances by one does not appear to diminish greatly as we move from 1 to 2 to 3 released ambulances.

Figure 2.11 illustrates the impact of requesting new ambulances ($n = 1, 2, 3$)

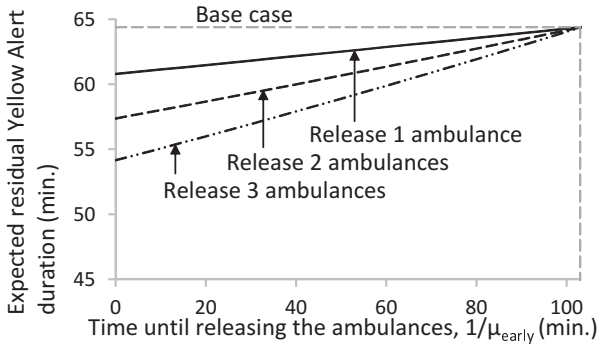
and the average time until the new ambulances are available ($1/\delta = 0, 5, ..., 100$ minutes). In contrast to the straight lines in Figure 2.10, the curves in Figure 11 are nonlinear, approaching the horizontal base-case line asymptotically. This happens because increasing the time to free the ED ambulances beyond the original value of $1/\mu$ results in performance that is *worse* than the base case, whereas increasing the average time until new ambulances are available causes the impact of requesting a new ambulance to approach zero, compared to the base case. We see clear evidence of diminishing marginal impact of requesting new ambulances, especially for the lost calls performance measure.

When we look closely at Figure 2.11(a) we see an unexpected pattern, whereby the residual Yellow Alert duration *increases* slightly when the average time until the requested ambulances become available decreases from roughly 3 minutes to 0. To understand why this happens, consider the following two opposing impacts of a request for new ambulances on the residual Yellow Alert duration: (1) It shortens the Yellow Alert duration because it ends when the number of busy ambulances drops to $k+n-1$ (rather than $k-1$), and (2) it increases the Yellow Alert duration because increasing the number of servers enables the system to serve additional customer(s), who would have been lost if the system did not have the additional n servers. When k' and the ambulance-arrival rate are large enough, Impact (2) is larger than Impact (1) and requesting new ambulances increases the expected residual Yellow Alert duration.

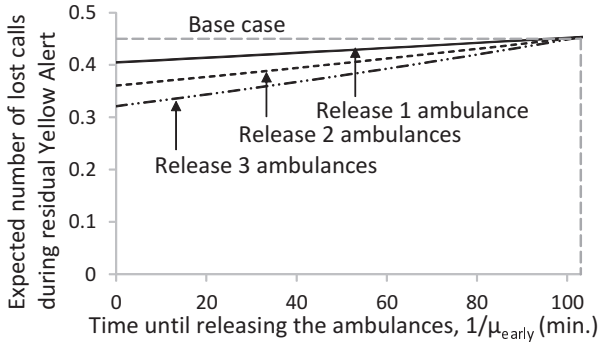
We summarize our observations from the sensitivity analyses, based on the results reported in Figures 2.10-2.11 and additional scenarios that are not reported here. The marginal impacts of μ_{new}, n, and $1/\delta$, holding all other parameters constant, appear to be as follows for the $M/M/c/c$ system:

- The expected residual Yellow Alert duration $\mathrm{E}\left(\tilde{B}_{kk'}\right)$ decreases when μ_{new} increases and decreases when n increases, but may or may not decrease when $1/\delta$ increases.

- The expected number of lost calls during the residual Yellow Alert duration



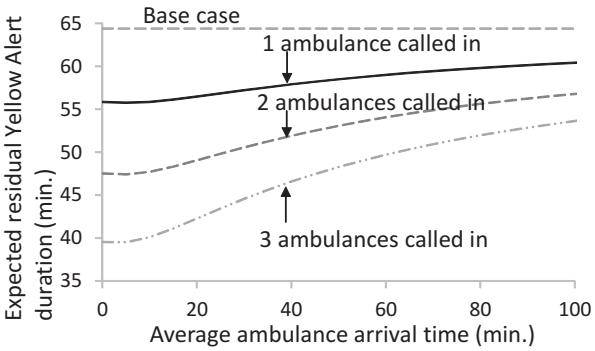
(a) Expected residual Yellow Alert duration.



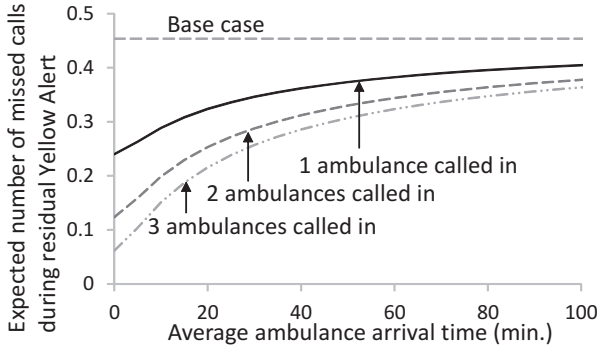
(b) Expected number of lost calls.

Figure 2.10: Sensitivity analysis: Released ambulances' numbers and times.
Note: In all scenarios, the arrival rate and the number of scheduled ambulances are equal to those of the base case: $\hat{\lambda}^{(4)} = 13.37$ ambulances per hour, $\hat{c}^{(4)} = 42$ ambulances. The base case service rate is $\hat{\mu}^{(4)} = 0.58$ patients per hour $\left(1/\hat{\mu}^{(4)} = 103 \text{ minutes}\right)$.

$\mathrm{E}\Big(H_{kk'}\Big)$ decreases when μ_{new} increases, n increases, and $1/\delta$ increases.



(a) Expected residual Yellow Alert duration.



(b) Expected number of lost calls.

Figure 2.11: Sensitivity analysis: Called-in ambulances' numbers and arrival rates.
Note: In all scenarios, the arrival rate and the number of scheduled ambulances are equal to those of the base case: $\hat{\lambda}^{(4)} = 13.37$ ambulances per hour, $\hat{c}^{(4)} = 42$ ambulances, $\hat{\mu}^{(4)} = 0.58$ patients per hour.

2.6.2 What Is the Best Combination of Actions?

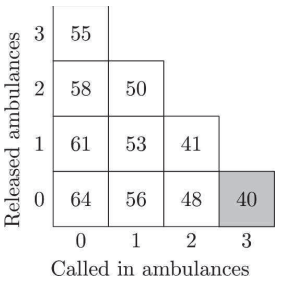
Suppose it costs c_1 to request an additional ambulance and c_2 to release an ambulance from an ED, respectively. If the total budget for taking corrective actions is b, then we find the best combination of corrective actions by solving the following optimization problems for the two performance measures:

$$\begin{array}{rl} \text{minimize} & z_1 = \mathrm{E}\left(\tilde{B}_{kk'}\right), \\ \text{subject to} & c_1 x_1 + c_2 x_2 \leq b, \\ & x_1,\, x_2 \text{ integer}, \end{array} \qquad \begin{array}{rl} \text{minimize} & z_2 = \mathrm{E}\Big(H_{kk'}\Big), \\ \text{subject to} & c_1 x_1 + c_2 x_2 \leq b, \\ & x_1,\, x_2 \text{ integer}. \end{array}$$

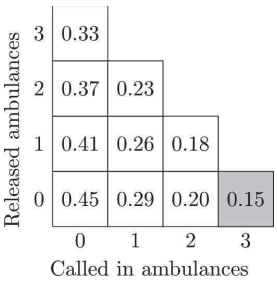
We do not attempt to quantify the cost coefficients and the budget in these optimization problems precisely. Rather, we view the optimization problems as a convenient way to provide information to dispatchers, by displaying the results of complete enumeration of all feasible combinations of actions, and identifying the optimal solution for each problem. Given that the calculations are near-instantaneous, dispatchers could experiment in real time with changing the cost coefficients and other input parameters.

Figure 2.12 shows complete enumeration results for a scenario where $k' = 40$, $c_1 = c_2 = 1$, $b = 3$, and $1/\delta = 1/\mu_{\text{early}} = 10$ minutes. As highlighted, spending the whole budget on requesting new ambulances is the best decision for both performance measures. To understand why this happens, suppose that we compare the impact of requesting one new ambulance with freeing one ambulance from an ED, assuming that the average time for the new ambulance to arrive and the time to free the ED ambulance are fixed to be equal to 10 minutes. From Figures 2.10-2.11, we see that requesting a new ambulance has greater impact on both the expected residual Yellow Alert duration (8 minutes vs. 3 minutes) and the expected number of lost calls (0.16 vs. 0.04).

In Figure 2.13, we show how the complete enumeration results change when we decrease the time to release ED ambulances from 10 to 0.001 minutes and increase

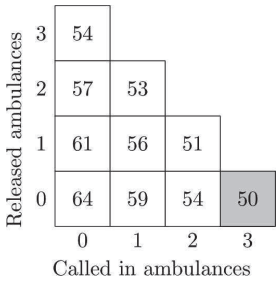


(a) E(res. Yellow Alert) in mins.

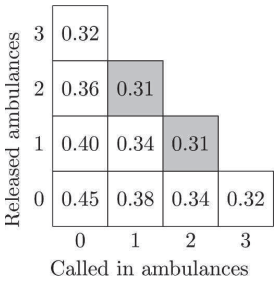


(b) Expected number of lost calls.

Figure 2.12: $k' = 40$, $c_1 = c_2 = 1$, $b = 3$, $1/\mu_{\text{early}} = 1/\delta = 10$ min.



(a) E(res. Yellow Alert) in mins.



(b) Expected number of lost calls.

Figure 2.13: $k' = 40$, $c_1 = c_2 = 1$, $b = 3$, $1/\mu_{\text{early}} = 0.001$ min., $1/\delta = 60$ min.

the time for new ambulances to become available from 10 to 60 minutes. In this scenario we see (1) a situation where the two performance measures lead to different optimal solutions and (2) that the optimal solution shifts to using a combination of freeing up ED ambulances and requesting new ambulances, at least for the lost calls performance measure.

2.7 Conclusion

This chapter provided an understanding of capacity shortage periods in mission critical systems like fire, police, and EMS. We focused on EMS systems and modeled these systems as Erlang loss systems ($M/G/c/c$) and showed that the expected duration of periods during which at least k out of c servers (ambulances) were busy were independent of the service time distribution shape beyond its mean. We obtained an easy-to-use recursion to calculate the expected duration of ambulance-

shortage periods. We validated our recursion against a year's worth of data from the Calgary EMS. By computing outputs separately for multiple time segments and weighing the time segments appropriately, we obtained a close match between the model outputs and the empirical outputs.

We calculated the LTs of the duration of server-shortage periods for the $M/G/c/c$ system and used these calculations to show that server-shortage periods had generalized hyper-Erlang and generalized hyperexponential distributions if service times had generalized hyperexponential and hyperexponential distributions, respectively. We discussed the monotonicity of the expected value, variance, and the squared coefficient of variation of the duration of server-shortage periods for the $M/M/c/c$ system.

We used the theory of absorbing Markov chains to predict the impacts of requesting new ambulances and of releasing ambulances in EDs. Our methods could be used for real-time decision support systems for EMS dispatchers—for example by showing how different combinations of requesting new ambulances and releasing ambulances in EDs impacted two performance measures: The expected remaining duration of a Yellow Alert and the number of lost calls during this residual duration. We showed that these two performance measures do, in some cases, lead to different optimal actions.

We did not consider costs associated with requesting new ambulances and expediting their service in our analyses as it was beyond the scope of this research. However, it would be interesting to consider these costs to study the impact of having different yellow alert thresholds on the system performance and to study policies for reversing the actions that are taken. Another direction for future work is considering the response time, period from the moment that a call is received until an ambulance unit arrives at the scene, as a performance measure. Then one can study the impact of taking any of the two actions on response time and investigate how this performance measure changes by the number of available ambulances. Note that when new ambulances are requested, travel times would tend to get shorter, but we ignore this fact in our model. Another limitation of current work is that we

ignore differences between the flexibility of the two actions that we study.

Endnote

1. Protocols that define when a medium intensity level or Yellow Alert is triggered sometimes include additional considerations besides the number of available units, such as "7 or fewer units ... sustained for 15 minutes" (from an overcapacity protocol for Edmonton Zone EMS and EDs published in 2011). We assume that alert periods are defined solely based on the number of available units, for simplicity.

CHAPTER 3

Modeling Queueing Systems With Abandonment And Priorities As Quasi-Birth-Death Processes

3.1 Introduction

Prolonged waiting times have turned to serious problems for emergency departments (ED) in most countries, and threaten patients' health and lives throughout the world (Hoot and Aronsky 2008, Higginson 2012). Studying almost 1 million ED visits during 2007 from non-federal, acute-care hospitals in California, Sun et al. (2013) report that crowded ED periods, measured by the number of diverted ambulances, can be associated with 300 inpatient deaths, 6,200 hospitalization days, and $17 million care costs. In addition, hospitals lose revenue when EDs are crowded as patients leave without being seen (LWBS) by a physician and ambulances get diverted to other hospitals. Pines et al. (2011) estimate that a hospital can increase its revenue by $10,000 to $13,000 per day, per 1-hour reduction in ED waiting times by capturing LWBS patients and diverted ambulances.

Timely access to ED physicians is a key factor in providing quality care for patients and increases the revenue for care provider institutes. Hospitals, however, compromise care quality and bear excessive costs as they struggle to provide adequate emergency care staff to handle the demand in a timely fashion. It is a difficult task for hospitals to constantly adjust their resources such that all patients receive timely access to care because the demand is non-homogeneous and highly variable. In their report prepared for the American Academy of Emergency Medicine, Eitel

Table 3.1: ESI suggested time lines for different acuity levels (Gilboy et al. 2011).

Acuity level	Suggested wait time
I: Resuscitation	Immediate
II: Emergent	$1-14$ minutes
III: Urgent	$15-60$ minutes
IV: Less urgent	$1-2$ hours
V: Non urgent	$2-24$ hours

et al. (2010) mention that the arrival rate of patients at EDs varies by time of the day, day of the week, and seasons, and not all patients have the same acuity level; some need immediate care while others can wait to see a physician.

An inherent complexity of an ED is that patients with different acuity levels need to be treated differently. Patients with life threatening conditions (like heart attack) have to be immediately seen by a physician, while others with less serious conditions (like stable abdominal pain) can be seen at a later time. When a patient arrives at an ED, a triage nurse assesses the patient's acuity level almost immediately, and assigns the patient to an acuity category. Guidelines for patient acuity assessment and patient wait times vary from one hospital to another even within a country. The Emergency Severity Index (ESI) is a widely used five-level ED triage guideline that assigns an ED patient to one of the following five categories: Resuscitation, emergent, urgent, less urgent and nonurgent (Gilboy et al. 2011), and suggests waiting time standards as shown in Table 3.1. Almost 60% of about 3,000 U.S. hospitals studied by McHugh et al. (2012) use ESI as their triage guideline.

In practice, however, patients may wait much longer than what guidelines recommend. The American College of Emergency Physicians (2013) reports that the average time for emergent patients to be seen by a physician was 37 minutes (more than twice as long as the recommended 14 minutes), according to national-level data. Carter et al. (2014) provide a systematic review on impacts of prolonged waiting times and report that previous studies find excessive waiting times are associated with higher LWBS rates. Research shows that LWBS patients do not necessarily belong to less and non urgent acuity levels (IV and V) (Rowe et al. 2006, Baker

et al. 1991).

Redesigning the ED flow process is considered as an innovative solution to mitigate prolonged waiting time issues. Sanchez et al. (2006) empirically demonstrate how an ED in the United States shortened waiting times and reduced the number of LWBS patients by creating a *fast track* for less acute patients through which these patients are seen by mid-level care providers (physician assistants and nurse practitioners) without occupying resources needed by more acute patients. Oredsson et al. (2011) provide a systematic literature review on studies that investigate impacts of different ED redesigns. Yildiz et al. (2015) investigate the impacts of an alternative ED flow redesign where the initial triage is provided by a physician (rather than a nurse) and patients with higher acuity levels are directed to a waiting room until a bed is assigned to them while less acute patients are treated by the triage physician.

Regardless of empirical findings that support creating fast tracks for EDs, some researchers are skeptical about their benefits. Lin et al. (2014), for example, argue that, unless additional resources are utilized, the fast track will improve performance measures for patients eligible for fast track and will worsen the performance measures for some other patients.

ED systems are so complex that it is not a trivial task to thoroughly understand the impact of fast track, or other ED flow redesigns, on performance measures. Viewed as a queueing system, an ED has multiple servers (care providers), multiple priority classes (acuity classes), and abandonment (LWBS patients). The analysis of multi-server queues with multiple classes of impatient customers is challenging, as we outline in the next section.

Organ transplantation systems are another example of systems that can be viewed as multi-server queues with multiple impatient priority classes. In these systems patients with different acuity levels are added to waiting lists to receive organs and might die (abandon the system) while waiting. Drekic et al. (2015) model and study an organ transplant problem as a single-server queue with two classes of impatient customers.

In this work, we use the performance evaluation problem for EDs as our moti-

vation and model multi-server queues with priorities and abandonment as level-dependent quasi-birth-and-death (LDQBD) processes. We specifically focus on queueing system with two priority classes and propose generic methods to calculate performance measures of these queueing systems. We propose an algorithm to analytically find a truncation level for the LDQBD process such that the probability mass in the truncated upper tail is guaranteed to be less than a pre-specified amount. We also propose another algorithm to numerically calculate stationary performance measures of the LDQBD process with desired accuracies.

3.2 Literature review

We focus our review of past work on three areas: 1) models of queues with priorities, 2) models of queues with priorities and abandonment, and 3) LDQBD processes.

Queues with priorities: Cobham (1954, 1955) calculates the expected waiting times of customers from different priority groups in $M/G/1$ and $M/M/c$ priority queues. Subsequent work has extended Cobham's work from expected waiting times to waiting time distribution (Davis 1966) and waiting time Laplace Transform (Kella and Yechiali 1985) when the priority groups have identical service time distributions. There are papers in the literature that study queueing systems with priority classes that have non-identical service time distributions. Miller (1981a) derive the steady state distribution of the $M/M/1$ system with two priority classes with different service rates, for both preemptive and non-preemptive settings and Bose (2013, Chapter 4) obtains mean performance measures (mean waiting time, for example) of $M/G/1$ queues with multiple classes of customers with different average service times, and obtain the steady state distribution for $M/G/1$ queues with two classes of customers with different service time distributions. Kleinrock (1964), Kleinrock and Finkelstein (1967), and recently Stanford et al. (2014) extended the literature of priority queues by studying queueing systems in which priorities of customers increase by time.

Queues with priorities and abandonment: There exists a large body of papers on queues with abandonment (Palm 1957, Barrer 1957, Garnett et al. 2002, Zeltyn and Mandelbaum 2005, for example); there is, however, a relatively small number of papers that investigate abandonment for priority queues as these systems are usually too complex to be analytically tractable. Choi et al. (2001) obtain the steady state probabilities for $M/M/1$ systems with two customer classes that have different service rates where the higher-class customers have a preemptive priority over the lower class ones, and customers from the higher class have a deterministically-distributed patience while others are indefinitely patient. Brandt and Brandt (2004) extend the findings of Choi et al. (2001) to cases where the patience of customers from the higher class are generally distributed, and derive the steady state probabilities and the waiting time distribution of customers from the lower class. Rozenshmidt (2008) extend the literature by studying multi-server queues. They calculate the expected waiting time of customers from any class for an $M/M/c$ queue with two classes of impatient customers, where the service and patience rates are the same for both classes. Iravani and Balcıoğlu (2008) and Sarhangian and Balcıoğlu (2013) study six different single-server and multi-server queueing systems with patient and impatient customers, where the service rate is the same across different customer classes; the priority is preemptive or non-preemptive and the impatience rate may change across the classes. Different steady state performance measures have been obtained. Jouini and Roubos (2014) consider an $M/M/c$ queue with two classes of customers when one class has a non-preemptive priority over the other. They assume both customer classes have the same service and impatience rates, and then obtain the Laplace Transforms of different waiting time performance measures under different service disciplines. Wang et al. (2015) claim to be the first researchers who analytically analyze a multi-server system with impatient customer classes and a preemptive priority policy when the service rate is different across the classes. The authors obtain a closed form generating function for the number of customers from the lower priority group in the system for a system with two servers; for systems with more than two servers, they obtain the moments of the number of lower-priority customers in the

system. It appears that there is no analytical work on waiting times of low-priority customers in multi-server queues with impatient customer classes when the service rates are different across customer classes.

Quasi-birth-and-death processes: A generalization of a birth-death process is called a quasi-birth-and-death (QBD) process when univariate state variables are extended to bivariate state variables with the first and second dimensions called the level and phase, respectively, such that one-step transitions from a state are restricted to states in the same level or levels above and below (no restriction in the phase dimension). QBD processes provide a powerful framework for formulating and computing performance measures for a variety of queueing and other stochastic systems (See Latouche and Ramaswami 1999, Chapter 1, for a wide range of examples). Drekic et al. (2015), Campello et al. (2013), Delasay et al. (2013) and Sun (2008) discuss different applications of QBD processes in the healthcare sector while Kawanishi (2008) and Zhang et al. (2011) use these processes to study problems in other service sectors. There are two main types of QBDs: level-independent (the transition rates are independent of the level) and level-dependent (the transition rates depend on the level). Level-independent QBDs are analogous to birth-death processes with a geometric tail, leading to closed-form expressions for various performance measures. These closed-form expressions typically involve the so-called rate matrix, which usually has to be computed numerically. There has been a great deal of research on properties and algorithms for level-independent QBD processes Neuts (1981), Latouche and Ramaswami (1999). Level-dependent QBDs (LDQBD) are analogous to birth-death processes with birth or death rates that do not stabilize as the value of the state variable increases, such as the birth-death processes for $M/M/\infty$ and $M/M/c+M$ (Erlang A) queues. Except for the $M/M/\infty$ special case, typically one cannot find closed-form expressions for performance measures for such systems. Latouche and Ramaswami (1999, Chapter 12) demonstrate that performance measures of an LDQBD can be expressed in terms of rate matrices that depend on the level. Kharoufeh (2011) describes two major complications for calculating the steady state probabilities for a generic LDQBD with an infinite state space

as: 1) the state space must be truncated in an appropriate manner, and 2) the rate matrices need to be computed efficiently. Bright and Taylor (1995), Baumann and Sandmann (2012), and Baumann and Sandmann (2013) numerically compute the rate matrices and steady state probability vectors for infinite state space LDQBDs. The main shortcomings of their methods are: 1) the state space truncation processes are heuristic, and 2) there are no error bounds on calculated steady state probabilities. We address both of these shortcomings in this chapter.

We demonstrate that multi-server queues with priorities and abandonment can be modeled in a natural way as LDQBD processes. We specifically focus on queues with 2 impatient customer classes that have different service and patience rates. We use Lyapunov analysis as in Dayar et al. (2011) and truncate the state space such that the stationary probability mass in the truncated upper tail of the state space is below some tolerance. As another analytical tool, we provide algorithms to calculate the steady state probabilities and performance measures with any desired accuracy. Our algorithm automatically truncates the state space such that the error tolerance is satisfied.

3.3 Models and Definitions

Consider a continuous-time Markov chain $\{X_t, t \in \mathbb{R}^+\}$ on the state space $\mathbb{S} = \{(\ell, h) : \ell \in \mathbb{Z}^+, h \in \mathbb{Y}\}$, where $\mathbb{Z}^+ = \{0, 1, ...\}$ and $\mathbb{Y} = \{0, 1, ..., p\}$, for a given positive integer p. The stochastic process depicted by this Markov chain is a QBD process if its transitions from state $(0, h)$ are restricted to states $(0, h')$ and $(1, h')$, and its transitions from state (ℓ, h), for $\ell \in \mathbb{Z}^{++}$, where $\mathbb{Z}^{++} = \mathbb{Z}^+ \backslash \{0\}$, are restricted to states $(\ell - 1, h'), (\ell, h')$, and $(\ell + 1, h')$, for $h, h' \in \mathbb{Y}$ (Latouche and Ramaswami 1999, Chapter 6). See Table B.1 for a list of notations in this chapter.

The first coordinate of state (ℓ, h) is known as its *level*, and the second coordinate is known as its *phase*. We focus on positive recurrent QBD processes with infinite number of levels and a finite fixed number of phases.

The infinitesimal generator matrix $\boldsymbol{Q}$ of such a QBD Markov chain is block

tridiagonal:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{A}_1^{(0)} & \boldsymbol{A}_0^{(0)} & & & \\ \boldsymbol{A}_2^{(1)} & \boldsymbol{A}_1^{(1)} & \boldsymbol{A}_0^{(1)} & & \\ & \boldsymbol{A}_2^{(2)} & \boldsymbol{A}_1^{(2)} & \boldsymbol{A}_0^{(2)} & \\ & & \boldsymbol{A}_2^{(3)} & \boldsymbol{A}_1^{(3)} & \ddots \\ & & & \ddots & \ddots \end{pmatrix}, \tag{3.1}$$

where the matrix blocks $\boldsymbol{A}_0^{(\ell)}$ and $\boldsymbol{A}_1^{(\ell)}$, $\ell \in \mathbb{Z}^+$, denote the transition rates from level ℓ to levels $\ell+1$ and ℓ, respectively, and the matrix blocks $\boldsymbol{A}_2^{(\ell)}$, $\ell \in \mathbb{Z}^{++}$, denote the transition rates from level ℓ to level $\ell - 1$. All of the matrix blocks in $\boldsymbol{Q}$ are of size $(p+1) \times (p+1)$.

Example 3.1. *Erlang A model.* Assume customers arrive following a Poisson process with rate λ to a queueing system with c servers. The customers' service and patience times are independently and identically distributed exponential with rates μ and γ, respectively. This is the standard Erlang A model. We include it as a test case, where we can use Algorithm B.1 from Ingolfsson and Tang (2012) to compute the stationary distribution in order to illustrate our Lyapunov analysis in Section 3.6 in a simple setting.

To model the system in Example 3.1 as a QBD process, we consider the number of customers ℓ in the system as the level. In this QBD process, $p = 0$ as there is only one state in each level. When the system is in state $(\ell, 0)$, either transition type 1 occurs and the system moves to state $(\ell + 1, 0)$ after a customer arrival, or transition type 2 occurs and the system moves to state $(\ell - 1, 0)$ (if $\ell > 0$). We use $s = \min(c, \ell)$ and $q = \max(\ell - c, 0)$ to denote the number of customers in service and in the queue, respectively. Table 3.2 shows the two transitions types and their associated rates where the function $\alpha_j(\ell, h)$ denotes the transition rate from state (ℓ, h) in the direction of transition class $j \in \{1, ..., J\}$. The infinitesimal generator matrix has block matrices of size 1×1:

Table 3.2: Transitions of an Erlang A system from state $(\ell, 0)$.

j	Transition type	To	Rate ($\alpha_j(\ell, 0)$)
1	Customer arrival	$(\ell + 1, 0)$	λ
2	Customer departure	$(\ell - 1, 0)$	$s\mu + q\gamma$

$$\boldsymbol{Q} = \begin{pmatrix} -\lambda & \lambda & & & & & & \\ \mu & -(\lambda+\mu) & \lambda & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & & c\mu & -(\lambda + c\mu) & \lambda & & \\ & & & & c\mu + \gamma & -(\lambda + c\mu + \gamma) & \lambda & \\ & & & & & \ddots & \ddots & \ddots \end{pmatrix}. \quad (3.2)$$

Example 3.2. *A single-server queue with two priority classes.* Assume that there are two customer classes, each with an independent Poisson arrival process with rate λ_i, independent exponentially distributed service times with rate μ_i, and independent exponentially distributed patience times with rate γ_i, $i = 1, 2$. Class-1 customers have preemptive priority over Class-2 customers: Class-2 customers can receive service only if there is no Class-1 customer in the system, and if a Class-2 customer is receiving service when a Class-1 customer arrives, then the Class-2 customer immediately returns to the queue freeing the server for the arriving Class-1 customer.

To approximate the system in Example 3.2 as a QBD process, we need to decide how to map the Class-1 and Class-2 customer counts to the level and the phase. Given that Class-1 customers receive priority, truncating the maximum number of Class-1 customers as finite should result in less error, and therefore we associate the Class-1 customers with the phase and the Class-2 customers with the level. Because of the preemptive policy, the Class-1 customer sub-system evolves independently of Class-2 customers. Therefore, if there was no limit p on the count of Class-1 customers, then the Class-1 sub-system would reduce to an Erlang A queue. This reduction suggests that a reasonable value for p can be determined by approximating

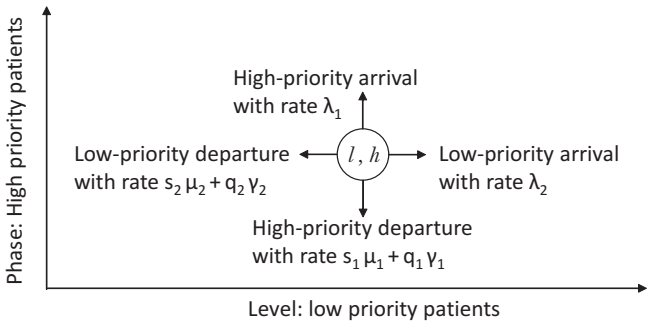


Figure 3.1: Transitions of a 2-priority multi-server queue from state (ℓ, h).

Table 3.3: Transitions of a 2-priority 1-server queue from state (ℓ, h).

j	Transition type	To	Rate ($\alpha_j(\ell, h)$)
1	Class-1 arrival	$(\ell, h+1)$	$\mathbb{1}_{<p}(h)\lambda_1$
2	Class-2 arrival	$(\ell+1, h)$	λ_2
3	Class-1 departure	$(\ell, h-1)$	$s_1\mu_1 + q_1\gamma_1$
4	Class-2 departure	$(\ell-1, h)$	$s_2\mu_2 + q_2\gamma_2$

the Class-1 sub-system with an Erlang A queue and choosing a sufficiently large p so that the probability of the number in the Erlang A system being greater than p is less than a small value ϵ_h, where ϵ_h is a parameter given by the user denoting desired accuracy. Based on the algorithm developed by Ingolfsson and Tang (2012) to compute performance measures for birth-death processes, we propose Algorithm B.1 to compute an appropriate value for p.

Figure 3.1 illustrates possible transitions of the system from state (ℓ, h). When the system is in state (ℓ, h), we use the binary variables $s_1 = \min\{h, 1\}$ to denote the number of Class-1 customers receiving service, $e = 1 - s_1$ to denote the number of servers that are not busy with Class-1 customers, and $s_2 = \min\{\ell, e\}$ to denote the number of Class-2 customers receiving service. We use $q_1 = h - s_1$ and $q_2 = \ell - s_2$ to denote the number of Class-1 and Class-2 customers in the queues, respectively. Possible transitions of this system are described in Table 3.3 where the indicator function $\mathbb{1}_A(x)$ equals 1 when x satisfies condition A, and equals 0 otherwise.

When $p = 2$, the infinitesimal generator matrix blocks are:

$$\boldsymbol{A}_0^{(\ell)} = \begin{pmatrix} \lambda_2 & & \\ & \lambda_2 & \\ & & \lambda_2 \end{pmatrix}, \ \boldsymbol{A}_1^{(\ell)} = \begin{pmatrix} * & \lambda_1 & \\ \mu_1 & * & \lambda_1 \\ & \mu_1 + \gamma_1 & * \end{pmatrix}, \ l \in \mathbb{Z}^+,$$

$$\boldsymbol{A}_2^{(\ell)} = \begin{pmatrix} \mu_2 + q_2\gamma_2 & & \\ & \ell\gamma_2 & \\ & & \ell\gamma_2 \end{pmatrix}, l \in \mathbb{Z}^{++}.$$

We use "*" for a generic diagonal element, whose value is chosen such that $\boldsymbol{Q}$ has zero row sums.

Example 3.3. *A multi-server queue with two priority classes.* We extend Example 3.2 by assuming that the system has $s > 1$ servers.

As we did in Example 3.2, we consider the number of Class-2 customers as the level ℓ, and the number of Class-1 customers as the phase h. We approximate the Class-1 sub-system with an Erlang A queue and choose p large enough such that the probability of the number in the Erlang A system being greater than p is less than ϵ_h. We determine p in the same fashion as we did for Example 3.2.

The possible transitions for Example 3.3 system is similar to those of Example 3.2 in Figure 3.1 and Table 3.3; however, the parameters s_1, s_2, q_1, and q_2 have to be redefined as follows. When the system is in state (ℓ, h), we use $s_1 = \min\{h, s\}$ to denote the number of Class-1 customers receiving service, $e = s - s_1$ to denote the number of servers that are not busy with Class-1 customers, and $s_2 = \min\{\ell, e\}$ to denote the number of Class-2 customers receiving service. We use $q_1 = h - s_1$ and $q_2 = \ell - s_2$ to denote the number of Class-1 and Class-2 customers in the queues, respectively.

The infinitesimal generator matrix blocks are:

$$\boldsymbol{A}_0^{(\ell)} = \begin{pmatrix} \lambda_2 & & \\ & \ddots & \\ & & \lambda_2 \end{pmatrix}, \ \boldsymbol{A}_2^{(\ell)} = \begin{pmatrix} s_2\mu_2 + q_2\gamma_2 & & \\ & \ddots & \\ & & s_2\mu_2 + q_2\gamma_2 \end{pmatrix}, \quad \ell \in \mathbb{Z}^{++},$$

$$\boldsymbol{A}_1^{(\ell)} = \begin{pmatrix} * & \lambda_1 & & & & \\ s_1\mu_1 + q_1\gamma_1 & * & \lambda_1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & s_1\mu_1 + q_1\gamma_1 & * & \lambda_1 & \\ & & & s_1\mu_1 + q_1\gamma_1 & * & \end{pmatrix}, \quad \ell \in \mathbb{Z}^{+}.$$

3.4 Review of LDQBD Theory and Algorithms

The matrices $\boldsymbol{G}^{(\ell)}$ and $\boldsymbol{R}^{(\ell)}$, for $\ell \in \mathbb{Z}^+$, play an important role in analyzing QBD processes (Latouche and Ramaswami 1999). The (i,j)th element of $\boldsymbol{G}^{(\ell)}$ is the probability of visiting state (ℓ, j) in the first visit to level ℓ when the process starts from state $(\ell+1, i)$, and the (i,j)th element of $\boldsymbol{R}^{(\ell)}$ is the average sojourn time in state $(\ell+1, j)$ before the first return to level ℓ per unit time in state (ℓ, i) provided that the system started at (ℓ, i).

Matrix $\boldsymbol{G}^{(\ell)}$ is a stochastic matrix and is related to matrix $\boldsymbol{R}^{(\ell)}$ through the following equations (Latouche and Ramaswami 1999), for $\ell \in \mathbb{Z}^+$:

$$\boldsymbol{G}^{(\ell)} = \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\boldsymbol{G}^{(\ell+1)}\right)^{-1} \boldsymbol{A}_2^{(\ell+1)} \tag{3.3a}$$

$$\boldsymbol{R}^{(\ell)} = \boldsymbol{A}_0^{(\ell)} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\boldsymbol{G}^{(\ell+1)}\right)^{-1} \tag{3.3b}$$

$$= \boldsymbol{A}_0^{(\ell)} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{R}^{(\ell+1)}\boldsymbol{A}_2^{(\ell+2)}\right)^{-1} \tag{3.3c}$$

We use the row-vector $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, ...)$ to denote the stationary distribution of the QBD process, where $\boldsymbol{\pi}_\ell = (\pi_{\ell,0}, ..., \pi_{\ell,p})$ is a row vector and $\pi_{\ell,h}$ denotes the steady-state probability of being in state (ℓ, h), $\ell \in \mathbb{Z}^+$ and $h \in \mathbb{Y}$. The rate matrices provide a way to compute the vectors π_ℓ recursively (Latouche and Ramaswami

1999):

$$\boldsymbol{\pi}_{\ell+1} = \boldsymbol{\pi}_{\ell}\boldsymbol{R}^{(\ell)}, \quad \ell \in \mathbb{Z}^{+}. \tag{3.4a}$$

The steady state probability vector is the solution to $\boldsymbol{\pi Q} = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is a column vector of ones of appropriate size. Following standard QBD analysis, using (3.4a) for $\ell = 0$, $\boldsymbol{\pi Q} = \mathbf{0}$ leads to an equation for π_0:

$$\boldsymbol{\pi}_0 \left(\boldsymbol{A}_1^{(0)} + \boldsymbol{R}^{(0)}\boldsymbol{A}_2^{(1)} \right) = \mathbf{0}. \tag{3.4b}$$

Using (3.4a) to express $\boldsymbol{\pi}_\ell$ in terms of $\boldsymbol{\pi}_0$, the normalization condition becomes:

$$\left(\sum_{\ell=0}^{\infty} \boldsymbol{\pi}_\ell \right) \mathbf{1} = \boldsymbol{\pi}_0(\boldsymbol{I} + \boldsymbol{R}^{(0)} + \boldsymbol{R}^{(0)}\boldsymbol{R}^{(1)} + \cdots)\mathbf{1} = 1. \tag{3.4c}$$

where $\boldsymbol{I}$ is an identity matrix of appropriate size.

Algorithm 3.1 outlines Baumann and Sandmann's (2013) first algorithm to compute an estimate $\widehat{\boldsymbol{\pi}}$ of the stationary distribution $\boldsymbol{\pi}$ by truncating the system at level $n < \infty$.

This algorithm has two limitations. It is not clear how to choose the truncation level n, and it is not clear how the truncation impacts the accuracy of the estimate $\widehat{\boldsymbol{\pi}}$ of the stationary distribution. We address these limitations in the remainder of this chapter.

Algorithm3.1: Baumann and Sandmann (2013) algorithm for $\boldsymbol{\pi}_\ell$ estimates.
Input: Truncation level n,
Initialization: $\widehat{\boldsymbol{R}}^{(n)} = \mathbf{0}$,
For $\ell = n-1, ..., 0$, compute $\boldsymbol{R}^{(\ell)} = \boldsymbol{A}_0^{(\ell)} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{R}^{(\ell+1)}\boldsymbol{A}_2^{(\ell+2)} \right)^{-1}$,
Solve $\widehat{\boldsymbol{x}}_0 \left(\boldsymbol{A}_1^{(0)} + \widehat{\boldsymbol{R}}^{(0)}\boldsymbol{A}_2^{(1)} \right) = \mathbf{0}$ to obtain $\widehat{\boldsymbol{x}}_0$,
For $\ell = 0, ..., n-1$, compute $\widehat{\boldsymbol{x}}_{\ell+1} = \widehat{\boldsymbol{x}}_\ell \widehat{\boldsymbol{R}}^{(\ell)}$,
Normalizing factor: $\widehat{c} = \left(\sum_{\ell=0}^{n} \widehat{\boldsymbol{x}}_\ell\right) \mathbf{1}$,
$\widehat{\boldsymbol{\pi}}_\ell = \widehat{\boldsymbol{x}}_\ell / \widehat{c}$, for $\ell = \{0, ..., n\}$.

3.5 Approaches to Determine Truncation Levels and Error Bounds

We propose two methods to provide bounds on the accuracy of performance measures that are calculated with Algorithm 3.1: (1) Inspired by Dayar et al. (2011), we use Lyapunov theory to compute a truncation level so that the truncated upper-tail state space is guaranteed to contain at most a pre-specified proportion of the stationary probability mass. Our Lyapunov analysis is separate and independent of Algorithm 3.1 and provides a method to calculate the parameter n for Algorithm 3.1. (2) We extend Algorithm 3.1 in such a way that our proposed algorithm endogenously determines the truncation level n and guarantees a pre-specified tolerance on the elements in the vector $\pi_{\ell'}$ of probabilities for a given level $\ell' \in \mathbb{Z}^+$.

3.6 Lyapunov Analysis to Determine Truncation Level

Lyapunov functions are used to analyze the stability of dynamical systems (Luenberger 1979), including Markov chains (Meyn and Tweedie 1993). Lyapunov functions can also be used to determine truncation limits for Markov chains such that the stationary probability mass in the truncated upper-tail state space is below some tolerance as illustrated by Dayar et al. (2011).

We will use the equilibrium point of a mean-field approximation in order to derive Lyapunov functions. A mean-field approximation for a continuous time Markov chain replaces discrete and random transitions with continuous and deterministic transitions (Izquierdo et al. 2011). The original process X_t transitions to $X_t + (\delta_{\ell,j}, \delta_{h,j})$ at exponential rate $\alpha_j(X_j)$ for $j = 1, \ldots, J$, where $(\delta_{\ell,j}, \delta_{h,j})$ is the vector of changes to the state variables for Transition j. In contrast, in the mean-field approximation, $\tilde{X}_t$ changes to $\tilde{X}_t + (\tilde{\delta}_{\ell,j}, \tilde{\delta}_{h,j}) \times \Delta t + o(\Delta t)$ as time changes from t to $t + \Delta t$, where the deterministic rates of change are computed as follows:

$$\tilde{\delta}_{\ell,j} = \sum_{j=1}^{J} \alpha_j(\ell, h)\delta_{\ell,j} \text{ and } \tilde{\delta}_{h,j} = \sum_{j=1}^{J} \alpha_j(\ell, h)\delta_{h,j}. \tag{3.5}$$

The equilibrium point (ℓ^*, h^*) of the mean-field approximation of a Markov chain

is the point for which the expected outgoing and incoming rates in each state variable direction are equal. That is,

$$\sum_{j\in\mathbb{I}^{(1,0)}_{(\ell^*,h^*)}} \alpha_j(\ell^*,h^*) = \sum_{j\in\mathbb{I}^{(-1,0)}_{(\ell^*,h^*)}} \alpha_j(\ell^*,h^*), \tag{3.6}$$

$$\sum_{j\in\mathbb{I}^{(0,1)}_{(\ell^*,h^*)}} \alpha_j(\ell^*,h^*) = \sum_{j\in\mathbb{I}^{(0,-1)}_{(\ell^*,h^*)}} \alpha_j(\ell^*,h^*), \tag{3.7}$$

where $\mathbb{I}^{(i,j)}_{(\ell,h)}$ denote the set of transitions that take the state (ℓ,h) to state $(\ell+i,h+j)$.

We introduce a Lyapunov function $g(\ell,h) : \mathbb{S} \to \mathbb{R}^+$ for queueing systems with two impatient customer classes and use this function together with results in Dayar et al. (2011, Section 3) to calculate the truncation level n so that $(\sum_{\ell=0}^{n}\pi_\ell)\,\mathbf{1} \geq 1-\epsilon_\ell$, or equivalently $\left(\sum_{\ell=n+1}^{\infty}\pi_\ell\right)\mathbf{1} < \epsilon_\ell$, for a given $\epsilon_\ell > 0$ (ℓ in ϵ_ℓ stands for the level). This truncation level can be used in Step 1 of Algorithm 3.1, instead of using heuristics to obtain n.

As Dayar et al. (2011) discuss, $g(\ell,h)$ is a Lyapunov function if a finite set $\mathbb{C}$ and a real number $\rho > 0$ exist such that the function g and its drift, $d_g(\ell,h)$, the average change in the value of g after a transition, satisfy the following conditions:

1. The drift is negative outside $\mathbb{C}$: $d_g(\ell,h) \leq -\rho$, $\forall(\ell,h) \in \mathbb{S}\backslash\mathbb{C}$,
2. The drift is finite within $\mathbb{C}$: $d_g(\ell,h) < \infty$, $\forall(\ell,h) \in \mathbb{C}$,
3. The sublevel sets of the Lyapunov function are finite: $\{(\ell,h)|g(\ell,h) \leq r\}$ is finite for all $r < \infty$.

The drift of $g(\ell,h)$ is calculated as (Dayar et al. 2011, Equation (4)):

$$d_g(\ell,h) = \sum_{j=1}^{J} \alpha_j(\ell,h)(g(\ell+\delta_{\ell,j}, h+\delta_{h,j}) - g(\ell,h)). \tag{3.8}$$

If $g(\ell,h)$ is a Lyapunov function, then one can set $\rho = z/\epsilon_\ell - z$, where $z = \sup\{d_g(\ell,h)|(\ell,h) \in \mathbb{S}\}$, and form set $\mathbb{C} = \{(\ell,h) \in \mathbb{S}|d_g(\ell,h) > -\rho\}$. Dayar et al. (2011, Page 1012) show that $\left(\sum_{(\ell,h)\in\mathbb{C}}\pi_{(\ell,h)}\right) \geq 1-\epsilon_\ell$. That is, more than $1-\epsilon_\ell$ percent of the probability resides in Level $n = \max\{\ell|(\ell,h) \in \mathbb{C}\}$ and below.

We use the Euclidean distance from the equilibrium point (ℓ^*, h^*) of the mean-field approximation of the QBD process as our Lyapunov function. We first illustrate our Lyapunov analysis by applying it to an Erlang A system (Example 3.1), and then we apply our method to a more complicated multiple-server systems with two impatient customer classes (Example 3.3). We skip Example 3.2 because it is a special case of Example 3.3, with $c = 1$.

Example 3.4. *Example 3.1 continued.*

We show that the following is a Lyapunov function for the Erlang A system:

$$g(\ell, 0) = (\ell - \ell^*)^2, \text{ for } \ell \geq 0, \tag{3.9}$$

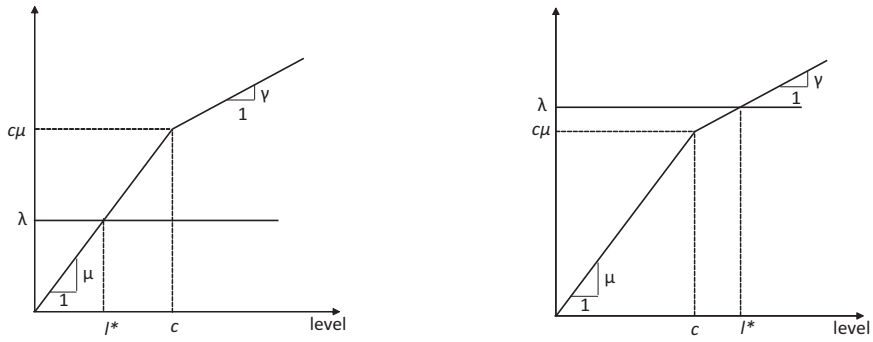
where ℓ^* satisfies $\lambda = s(\ell)\mu + q(\ell)\gamma = \min(c, \ell)\mu + \max(\ell - c, 0)\gamma$; this equation is obtained by combining (3.6) and Table 3.2 formulas. The mean-field equilibrium point occurs where the Erlang A birth rate (λ) equals the death rate ($s\mu + q\gamma$). The death rate is piece-wise linear in ℓ, with slope μ for $\ell \in [0, c)$ and slope γ for $\ell \in [c, \infty)$, as shown in Figure 3.2. Since the death rate is strictly increasing, there is a unique level at which the birth rate equals the death rate, and this is the mean-field equilibrium point ℓ^*. Consideration of the two cases where λ is either above or below the death rate $c\mu$ when $\ell = c$ leads to the following expression for ℓ^*:

$$\ell^* = \begin{cases} \lambda/\mu & \text{if } \lambda/\mu < c \\ (\lambda - (\mu - \gamma)c)/\gamma & \text{if } \lambda/\mu \geq c. \end{cases} \tag{3.10}$$

We apply (3.8) and obtain the drift of (3.9) in terms of the system parameters:

$$\underset{\ell \geq 0}{d_g(\ell, 0)} = \begin{cases} 2(\ell - \ell^*)(\lambda - \mu\ell) + \lambda + \mu\ell & \text{if } \lambda/\mu < c \\ 2(\ell - \ell^*)(\lambda - c\mu - \gamma\ell + c\gamma) + \lambda + c\mu + \gamma\ell - c\gamma & \text{if } \lambda/\mu \geq c. \end{cases} \tag{3.11}$$

where ℓ^* is obtained using (3.10). The first expression is for the case where the offered load, λ/μ, is less than the number of servers, and the second expression



(a) $\lambda/\mu < c$ (b) $\lambda/\mu \geq c$

Figure 3.2: The value of ℓ^* depends on the relative magnitude of $c\mu$ and λ.

is for the case where the offered load exceeds the number of servers. The system remains stable even in the second case, because of customer abandonment.

One can confirm that both functions in (3.11) are concave quadratic functions of ℓ, and for any $\ell_0 \in \mathbb{Z}^+$ that is larger than (3.11) roots, we have:

- $d_g(\ell_0, 0) < 0$
- $d_g(\ell_0, 0) \leq d_g(\ell, 0)$ for $0 \leq \ell \leq \ell_0$
- $d_g(\ell, 0)$ decreases when ℓ increases for $\ell > \ell_0$

Therefore, if $\mathbb{C} = \{(\ell, 0) \in \mathbb{S} | \ell \leq \ell_0\}$ and $\rho = -d_g(\ell_0, 0)$, then $d_g(\ell, 0) \leq -\rho$, $\forall (\ell, 0) \in \mathbb{S} \backslash \mathbb{C}$ (Condition 1 holds). Furthermore, $d_g(\ell, 0) < \infty$, $\forall (\ell, 0) \in \mathbb{C}$ (Condition 2 holds), and the number of states $(\ell, 0) \in \mathbb{S}$ that satisfy $\ell \leq \ell^* + \sqrt{r}$ is finite for all $r < \infty$ (Condition 3 holds).

We assume $c \leq \lambda/\mu$—one can repeat the analysis for $\lambda/\mu \leq c$ in the same fashion as we do here. The maximum of (3.11) happens at $\ell_{\max} = 0.25 + \lambda/\mu$, and $z = d_g(\ell_{\max}, 0) = 2(\lambda + 0.0625\mu)$. Then we set $\rho = z/\epsilon_\ell - z$ and form $\mathbb{C} = \{(\ell, 0) \in \mathbb{S} | d_g(\ell, 0) > -\rho\}$ to obtain $n = \max\{\ell | (\ell, 0) \in \mathbb{C}\}$.

Table 3.4 shows states that include more than 99% ($\epsilon_\ell = 0.01$) of the steady state probabilities for different scenarios of Erlang A systems when $\lambda = c = 10$ and $\gamma = 0.1$. System 1 has a low offered load per server of $\lambda/(c\mu) = 0.2$. We decrease the service rate and keep all other parameters unchanged, resulting in Systems 2 and 3

Table 3.4: Lyapunov analysis outputs for Example 3.4.

Parameter/ Output	System 1	System 2	System 3
μ	5	1	0.1
ℓ^*	2	10	100
ρ	2041.87	1981.24	1981.24
$\mathbb{C}$	$\{0, ..., 28\}$	$\{0, ..., 111\}$	$\{10, ..., 201\}$
n	28	111	201
n'	11	36	125
$\Pr(\ell > n)$	0.8734×10^{-25}	0.1192×10^{-21}	0.7856×10^{-22}

with offered loads per server of 1 and 10, respectively. As expected, the probability mass moves towards higher states with decreasing service rate.

As Dayar et al. (2011) discuss, the Lyapunov analysis may provide overly conservative ranges for a given tolerance ϵ_ℓ. We used Algorithm B.1 to calculate the smallest truncation level n' that satisfies $\left(\sum_{\ell=0}^{n'} \pi_\ell\right) \mathbf{1} \geq 0.99$, and we used Algorithm B.2 to compute the upper-tail probability above the Lyapunov truncation level n. We see in Table 3.4 that the Lyapunov truncation levels are indeed much larger than they need to be for these systems, providing error tolerances of smaller than 10^{-20} instead of the desired $\epsilon_\ell = 0.01$.

Example 3.5. *Example 3.3 continued.*

We show that the following is a Lyapunov function for multiple-server queueing systems with two impatient customer classes:

$$g(\ell, h) = (\ell - \ell^*)^2 + (h - h^*)^2, \text{ for } \ell, h \geq 0, \tag{3.12}$$

where ℓ^* and h^* satisfy $\mathbb{1}_{<p}(h)\lambda_1 = s_1(h)\mu_1 + q_1(h)\gamma_1$ and $\lambda_2 = s_2(\ell, h)\mu_2 + q_2(\ell, h)\gamma_2$; these equations are obtained by combining (3.6), (3.7) and Table 3.3 formulas. That

is,

$$h^* = \begin{cases} \lambda_1/\mu_1 & \text{if } c > \lambda_1/\mu_1, \\ (\lambda_1 + (\gamma_1 - \mu_1)c)/\gamma_1 & \text{if } c \leq \lambda_1/\mu_1, \end{cases} \tag{3.13}$$

$$\ell^* = \begin{cases} \lambda_2/\mu_2 & \text{if } c > \lambda_1/\mu_1 + \lambda_2/\mu_2, \\ (\lambda_2 + (\gamma_2 - \mu_2)(c - \lambda_1/\mu_1))/\gamma_2 & \text{if } \lambda_1/\mu_1 < c \leq \lambda_1/\mu_1 + \lambda_2/\mu_2, \\ \lambda_2/\gamma_2 & \text{if } c \leq \lambda_1/\mu_1 \end{cases} \tag{3.14}$$

The first expression of (3.13) is for the case where the offered load by Class-1 customers, λ_1/μ_1, is less than the number of servers, and the second expression is for the case where the Class-1 offered load exceeds the number of servers. For (3.14), the first expression is for the case where the system capacity is enough for both customer classes, the second expression is for the case where the system capacity is enough for the Class-1, but not enough for both classes, and the third expression is for the case where the system capacity is not enough even for Class-1 customers.

We apply (3.8) and obtain the drift of (3.12) in terms of the system parameters:

$$d_g(\ell, h) \underset{\ell,h\geq 0}{=} \begin{cases} (2\ell - 2\ell^*)(\lambda_2 - \mu_2\ell) + (2h - 2h^*)(\lambda_1 - \mu_1 h) + \\ \mu_2\ell + \mu_1 h + \lambda_1 + \lambda_2 & \text{if } c > \lambda_1/\mu_1 + \lambda_2/\mu_2, \\ (2\ell - 2\ell^*)(\lambda_2 - \mu_2(c-h) - \gamma_2(\ell - c + h)) + \\ (2h - 2h^*)(\lambda_1 - \mu_1 h) + \mu_2(c-h) + & \text{if } \lambda_1/\mu_1 < c \leq \\ \gamma_2(\ell - c + h) + \mu_1 h + \lambda_1 + \lambda_2 & \quad \lambda_1/\mu_1 + \lambda_2/\mu_2, \\ (2\ell - 2\ell^*)(\lambda_2 - \gamma_2\ell) + \\ (2h - 2h^*)(\lambda_1 - c - (h-c)\gamma_1) + \\ \gamma_2\ell + c + (h-c)\gamma_1 + \lambda_1 + \lambda_2 & \text{if } c \leq \lambda_1/\mu_1, \end{cases} \tag{3.15}$$

where ℓ^* and h^* are obtained using (3.13) and (3.14). Note that the system capacity for Class-1 customers, p, is an input that is chosen sufficiently large such that 1) As

discussed in Example 3.3, the probability of having more than p Class-1 customers in associated Erlang A system is less than ϵ_h, and 2) $p \geq h^*$ to guarantee that $\mathbb{1}_{<p}(h)\lambda_1 = s_1\mu_1 + q_1\gamma_1$ has a solution.

Each of the three equations in (3.15) is the sum of a concave quadratic function of ℓ and a concave quadratic function of h. In the same fashion as in Example 3.4, one can confirm that (3.12) is a Lyapunov function.

We assume that there is enough capacity to cover the offered load from both customer classes, that is $c > \lambda_1/\mu_1 + \lambda_2/\mu_2$—one can repeat the analysis for other cases in the same fashion as we do here. The maximum of (3.15) happens at $\ell_{\max} = 0.25 + \lambda_2/\mu_2$ and $h_{\max} = 0.25 + \lambda_1/\mu_1$, and $z = d_g(\ell_{\max}, h_{\max}) = 2(0.0625(\mu_1 + \mu_2) + \lambda_1 + \lambda_2)$. Then we set $\rho = z/\epsilon_\ell - z$ and form $\mathbb{C} = \{(\ell, h) \in \mathbb{S} | d_g(\ell, h) > -\rho\}$.

We illustrate the analysis as follows. Class-1 patients have more complex complaints, therefore their service rate is smaller than that of Class-2 patients and Class-1 patients are less likely to abandon compared to Class-2 patients; following this rationale, we set $\mu_1 = 1$, $\mu_2 = 2$, $\gamma_1 = 0.1$, and $\gamma_2 = 1$. We assume that 10% of the patients are Class 1 and 90% are Class 2, that is, $\lambda_1 = 0.1\lambda$ and $\lambda_2 = 0.9\lambda$. The resulting offered load is $R = \lambda_1/\mu_1 + \lambda_2/\mu_2$. We vary the offered load by varying λ, and we set the number of servers c equal to the offered load. We choose values of λ that result in integer values for the offered load. For a given λ, we choose p such that the probability of of having more than p patients in the Erlang A system associated with Class-1 sub-system is less than $\epsilon_h = 10^{-6}$. We choose a very small ϵ_h to virtually cover all of the possibilities for the number of Class-1 customers in the system. We set $\epsilon_\ell = 0.01$ and find Level n below which 99% of the total probability resides. Table 3.5 shows Level n for different systems with varying offered load from 1 to 2 and 3. As expected, the probability mass moves towards higher states when the offered load increases.

Table 3.5: Lyapunov analysis for Example 3.5.

Parameter/Output	System 4	System 5	System 6
λ	1.8189	3.6364	4.9091
$R = c$	1	2	3
ℓ^*	0.8182	1.6364	2.4545
h^*	0.1818	0.3636	0.5454
ρ	376.5039	769.4775	1294.4943
$\mathbb{C}$	$\{(0,0),...,(16,7)\}$	$\{(0,0),...,(23,7)\}$	{(0,0),...,(31,8)}
n	16	23	31
p	7	7	8

3.7 Extension of Algorithm 3.1

We focus on the LDQBD model of multi-server queues with two classes of impatient customers and propose an algorithm to calculate a lower bound and an upper bound for the steady state probability vector of a given level $\ell' \in \mathbb{Z}^+$, $\underline{\boldsymbol{\pi}}_{\ell'}$ and $\overline{\boldsymbol{\pi}}_{\ell'}$, respectively, such that $\overline{\boldsymbol{\pi}}_{\ell'}\mathbf{1} - \underline{\boldsymbol{\pi}}_{\ell'}\mathbf{1} \leq \epsilon_{\ell'}$, for any desired error tolerance $\epsilon_{\ell'}$. Note that this condition guarantees that all of the probabilities $\pi_{\ell',h}, h = 0, \ldots, p$ at Level ℓ' are known to within the tolerance $\epsilon_{\ell'}$.

To recall, Algorithm 3.1 starts from a heuristically-chosen Level n and assigns $\boldsymbol{R}^{(n)} = \mathbf{0}$. The recursive formula (3.3c) is used to calculate the rate matrices for Levels $n-1, ..., 0$. At the end, the normalizing factor and all state variables are calculated. In contrast, our Algorithm 3.2 starts from Level 0 and updates the normalizing factor and bounds on $\boldsymbol{\pi}_{\ell'}$ at the end of each iteration using a procedure that is discussed later in this section. Our algorithm proceeds until a truncation level k has reached that guarantees the bounds on $\boldsymbol{\pi}_{\ell'}$ to be within the error tolerance.

From the list of input parameters, we use λ_1, λ_2, μ_1, μ_2, γ_1, γ_2, c, and p to build the transition matrix blocks $\boldsymbol{A}_0^{(\ell)}$, $\boldsymbol{A}_1^{(\ell)}$, and $\boldsymbol{A}_2^{(\ell)}$ for different system levels. We use τ to denote a level above which the rate matrices are element-wise decreasing; that is, $\boldsymbol{R}^{(\tau+i)} \leq \boldsymbol{R}^{(\tau+i+1)}$ for $i \in \mathbb{Z}^{++}$. Finding τ is still an open problem. Later, in our numerical examples, we discuss a heuristic method that we use to choose τ.

Iteration ℓ of Algorithm 3.2—Iteration 0 includes lines 2-16 and Iteration $\ell \geq 1$ is the ℓth loop in lines 17-27—commences by calling Algorithm 3.3 that calculates

a lower bound and an upper bound for the rate matrix $\boldsymbol{R}^{(\ell)}$, which relies on bounds that Algorithm 3.3 calculates for $\boldsymbol{G}^{(\ell)}$. In order to obtain the $\boldsymbol{G}^{(\ell)}$ bounds, we compute approximations to $\boldsymbol{G}^{(\ell)}$ based on the assumption that the system never visits levels at or above $\ell + m$. We provide m as an input for Algorithm 3.3. An advantage of having a large m is that we obtain a better estimation of $\boldsymbol{G}^{(\ell)}$; a disadvantage, on the other hand, is that it would be time consuming to obtain the $\boldsymbol{G}^{(\ell)}$ estimation and we might encounter round-off errors. The choice of parameter m is discussed later in numerical examples.

Iteration ℓ of Algorithm 3.2 continues by calculating a lower bound and an upper bound for the unnormalized stationary probability $\boldsymbol{x}_\ell$. When $\ell = 0$, we use variations of (3.4b), and for $\ell > 0$, we use variations of (3.4a) to obtain the bounds. We proceed by calculating a lower bound and an upper bound on the normalizing factor c. At the end, we calculate bounds on the stationary probability of level ℓ' and check whether the bounds satisfy our error tolerance.

To illustrate how the algorithm works and discuss the choice of parameters τ and m, we calculate bounds on $\boldsymbol{\pi}_0$ for System 4 of Table 3.5 when $\epsilon_{\ell'} = 0.01$ and then repeat our numerical experiments on variations of System 4 by increasing λ. For System 4, we set $\tau = \lceil \ell^* \rceil$ because when the system is in level $\ell > \ell^*$, the expected next move would be towards level $\ell - 1$. It may indicate that the system tends to spend more time in level $\ell - 1$ than it does in level ℓ. According to the interpretation of the elements of a rate matrix, it follows that $\boldsymbol{R}^{(\ell-1)} \geq \boldsymbol{R}^{(\ell)}$. We choose the parameter m by trial and error. When $m = 5$, regardless of the number of iterations, we cannot calculate $\boldsymbol{\pi}_0$ within the 0.01 error tolerance. When we set $m = 30$, because of round-off errors, some elements of the lower bound for rate matrices become larger than the associated upper bound elements. When we set $m = 15$, the algorithm finds the bounds on $\boldsymbol{\pi}_0$ after 5 iterations. Figure 3.3 shows how the bounds on the first 3 elements of $\boldsymbol{\pi}_0$ become tighter and the $Gap = (\overline{\boldsymbol{\pi}}_0 - \underline{\boldsymbol{\pi}}_0)\mathbf{1}$ decreases by increasing the truncation level from one iteration to the next.

Table 3.6 shows the bounds on $\pi_{0,0}$ of multi-server multi-class queues similar to

Table 3.6: Bounds on the stationary probability $\pi_{0,0}$ of Example 3.5.

Parameter/Output	System 4	System 7	System 8	System 9
λ	1.82	9.09	18.18	54.54
p	7	9	11	31
$R = c$	1	5	10	30
τ	1	5	9	25
m	15	30	40	60
Lower bound	0.3038	0.0051	0.0000	0.0000
Upper bound	0.3071	0.0091	0.0009	4.1401×10^{-8}
Truncation level	4	5	4	10
Time (seconds)	5	11	13	43

Table 3.5 systems with offered loads that vary between 1 and 30. By increasing the load, the probability mass shifts towards higher states and the probability of being in state $(0, 0)$ decreases.

3.8 Error Bounds for New Algorithms

Proofs of all propositions in this section are provided in Appendix B.2. There are two types of errors in the $\widehat{\boldsymbol{\pi}}_{\boldsymbol{\ell}}$ values as estimated in Algorithm 3.1: *Rate-matrix error* and *level-truncation error* defined below. Let $\boldsymbol{x}_\ell$, $\ell = 0, ..., n$, be the vectors that would be obtained if the true rate matrices $\boldsymbol{R}^{(0)}, ..., \boldsymbol{R}^{(n-1)}$ were used in Algorithm 3.1. Define:

$$\Delta \boldsymbol{x}_\ell \coloneqq \boldsymbol{x}_\ell - \widehat{\boldsymbol{x}}_\ell, \quad \ell = 0, ..., n. \tag{3.16}$$

Using (3.16), we define and decompose the normalizing constant c as:

$$c \coloneqq \left(\sum_{\ell=0}^{\infty} \boldsymbol{x}_\ell \right) \mathbf{1} = \left(\sum_{\ell=0}^{n} \boldsymbol{x}_\ell \right) \mathbf{1} + \left(\sum_{\ell=n+1}^{\infty} \boldsymbol{x}_\ell \right) \mathbf{1} \tag{3.17}$$

$$= \left(\sum_{\ell=0}^{n} \widehat{\boldsymbol{x}}_\ell \right) \mathbf{1} + \left(\sum_{\ell=0}^{n} \Delta \boldsymbol{x}_\ell \right) \mathbf{1} + \left(\sum_{\ell=n+1}^{\infty} \boldsymbol{x}_\ell \right) \mathbf{1}, \tag{3.18}$$

and, re-write $\boldsymbol{\pi}_\ell$ as:

$$\boldsymbol{\pi}_\ell = \frac{\boldsymbol{x}_\ell}{c} = \frac{\widehat{\boldsymbol{x}}_\ell + \Delta\boldsymbol{x}_\ell}{\left(\sum_{\ell''=0}^{n} \widehat{\boldsymbol{x}}_{\ell''}\right)\mathbf{1} + \left(\sum_{\ell''=0}^{n} \Delta\boldsymbol{x}_{\ell''}\right)\mathbf{1} + \left(\sum_{\ell''=n+1}^{\infty} \boldsymbol{x}_{\ell''}\right)\mathbf{1}} \tag{3.19}$$

$$= \frac{\widehat{\boldsymbol{x}}_\ell + \Delta\boldsymbol{x}_\ell}{\widehat{c} + \left(\sum_{\ell''=0}^{n} \Delta\boldsymbol{x}_{\ell''}\right)\mathbf{1} + \left(\sum_{\ell''=n+1}^{\infty} \boldsymbol{x}_{\ell''}\right)\mathbf{1}}. \tag{3.20}$$

When computing $\widehat{\pi}_\ell$ using Algorithm 3.1, we call the error caused by ignoring $\Delta\boldsymbol{x}_\ell$ and $\left(\sum_{\ell''=0}^{n} \Delta\boldsymbol{x}_{\ell''}\right)\mathbf{1}$ the rate-matrix error and the error caused by ignoring $\left(\sum_{\ell''=n+1}^{\infty} \boldsymbol{x}_{\ell''}\right)\mathbf{1}$ the level-truncation error.

We use the following notational convention in our developments. Let $\boldsymbol{X}$ be a generic $(p+1)\times(p+1)$ matrix of unknown non-negative entries. We use $\underline{\boldsymbol{X}}$ to denote a known element-wise lower bound of $\boldsymbol{X}$ and $\overline{\boldsymbol{X}}$ to denote a known element-wise upper bound of $\boldsymbol{X}$; that is $\underline{\boldsymbol{X}} \leq \boldsymbol{X} \leq \overline{\boldsymbol{X}}$, where we use inequalities to denote element-wise orderings. Let $\Delta\boldsymbol{X} = \boldsymbol{X} - \underline{\boldsymbol{X}}$ and $\overline{\Delta\boldsymbol{X}} = \overline{\boldsymbol{X}} - \underline{\boldsymbol{X}}$, which implies that $\Delta\boldsymbol{X} \leq \overline{\Delta\boldsymbol{X}}$. Although the error matrix $\Delta\boldsymbol{X}$ is unknown (because $\boldsymbol{X}$ is unknown), its upper bound $\overline{\Delta\boldsymbol{X}}$ is computable (because the lower and upper bounds are known). We use $(\boldsymbol{X})_{i,j}$ to denote the (i,j)th entry of $\boldsymbol{X}$.

We develop all material needed to construct Algorithms 3.2-3.3 in 5 steps. In the first step, we develop bounds for the $\boldsymbol{G}$ matrices; in the second step, we develop bounds for the $\boldsymbol{R}$ matrices; in the third step, we obtain bounds for $\boldsymbol{x}_0$; in the fourth step, we develop bounds for the unnormalized stationary probabilities of higher levels and develop bounds for the normalizing factor; and in the last step se develop bounds on the stationary probability vector of interest.

Step 1: Bounds for $\boldsymbol{G}^{(\ell)}$, $\ell \in \mathbb{Z}^{++}$.

As we do not need $\boldsymbol{G}^{(0)}$ in our calculations, we skip $\boldsymbol{G}^{(0)}$. Following (Phung-Duc et al. 2010), for $\ell \in \mathbb{Z}^{++}$ and $m \in \mathbb{Z}^{+}$, we define $\boldsymbol{G}_m^{(\ell)}$ and $\boldsymbol{R}_m^{(\ell)}$ as estimates of $\boldsymbol{G}^{(\ell)}$ and $\boldsymbol{R}^{(\ell)}$, such that the (i,j)th element of $\boldsymbol{G}_m^{(\ell)}$ is the probability of visiting state (ℓ, j) in the first visit to level ℓ provided that the process starts from state $(\ell+1, i)$ and never visits the level $\ell + m + 1$ and levels above. And the (i,j)th element of $\boldsymbol{R}_m^{(\ell)}$ is the average sojourn time in state $(\ell+1, j)$ before the first return to level

ℓ per unit time in state (ℓ, i) provided that the system started at (ℓ, i) and never visits the level $\ell + m + 1$ and levels above.

Some properties of the $\boldsymbol{G}_m^{(\ell)}$ and $\boldsymbol{R}_m^{(\ell)}$ matrices are as follows (Phung-Duc et al. 2010): For $\ell \in \mathbb{Z}^{++}$ and $m \in \mathbb{Z}^+$,

Property 3.1. Similar to equations (3.3a)-(3.3c), $\boldsymbol{G}_m^{(\ell)}$ and $\boldsymbol{R}_m^{(\ell)}$ are related as follows:

$$\boldsymbol{G}_m^{(\ell)} = \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\boldsymbol{G}_{m-1}^{(\ell+1)}\right)^{-1} \boldsymbol{A}_2^{(\ell+1)} \tag{3.21a}$$

$$\boldsymbol{R}_m^{(\ell)} = \boldsymbol{A}_0^{(\ell)} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\boldsymbol{G}_{m-1}^{(\ell+1)}\right)^{-1} \tag{3.21b}$$

$$= \boldsymbol{A}_0^{(\ell)} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{R}_{m-1}^{(\ell+1)}\boldsymbol{A}_2^{(\ell+2)}\right)^{-1}, \tag{3.21c}$$

Property 3.2. $\boldsymbol{G}_0^{(\ell)} = \boldsymbol{R}_0^{(\ell)} = \boldsymbol{0}$,

Property 3.3. $\boldsymbol{G}_m^{(\ell)} \leq \boldsymbol{G}_{m+1}^{(\ell)}$, $\boldsymbol{R}_m^{(\ell)} \leq \boldsymbol{R}_{m+1}^{(\ell)}$,

Property 3.4. $\lim_{m\to+\infty} \boldsymbol{G}_m^{(\ell)} = \boldsymbol{G}^{(\ell)}$, $\lim_{m\to+\infty} \boldsymbol{R}_m^{(\ell)} = \boldsymbol{R}^{(\ell)}$,

Property 3.5. $\boldsymbol{G}_m^{(\ell)}$ is a sub-stochastic matrix: All entries are non-negative, and each row sum is at most 1.

To calculate $\boldsymbol{G}_m^{(\ell)}$, we define $L_\ell(\boldsymbol{X}) \coloneqq \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\boldsymbol{X}\right)^{-1} \boldsymbol{A}_2^{\ell+1}$ and combine it with (3.21a) as:

$$\boldsymbol{G}_m^{(\ell)} = L_\ell\left(L_{\ell+1}\left(\ldots\left(L_{\ell+m-1}\left(\boldsymbol{G}_0^{(\ell+m)}\right)\right)\right)\right), \quad \forall \ell \in \mathbb{Z}^{++}, \forall m \in \mathbb{Z}^+, \tag{3.22}$$

where $\boldsymbol{G}_0^{(\ell+m)} = \boldsymbol{0}$ due to Property 3.2. Employing Properties 3.3-3.4, one can confirm that $\boldsymbol{G}_m^{(\ell)} \leq \boldsymbol{G}^{(\ell)} = \lim_{m\to+\infty} \boldsymbol{G}_m^{(\ell)}$. Therefore, $\underline{\boldsymbol{G}}^{(\ell)} \coloneqq \boldsymbol{G}_m^{(\ell)}$ is a lower bound for $\boldsymbol{G}^{(\ell)}$ and its error is $\Delta\boldsymbol{G}^{(\ell)} = \boldsymbol{G}^{(\ell)} - \underline{\boldsymbol{G}}^{(\ell)}$.

To construct an upper bound $\overline{\boldsymbol{G}}^{(\ell)}$ on $\boldsymbol{G}^{(\ell)}$, we use the sub-stochastic matrix $\underline{\boldsymbol{G}}^{(\ell)}$. Define $a_i{}^{(\ell)} \coloneqq 1 -$ row sum of the ith row of $\underline{\boldsymbol{G}}^{(\ell)}$, for $i \in \mathbb{Y}$. We obtain $\overline{\boldsymbol{G}}^{(\ell)}$ by adding $a_i{}^{(\ell)}$ to each element in the ith row of $\underline{\boldsymbol{G}}^{(\ell)}$. The error matrix

$\overline{\Delta \boldsymbol{G}}^{(\ell)} = \overline{\boldsymbol{G}}^{(\ell)} - \underline{\boldsymbol{G}}^{(\ell)}$ is an upper bound on $\Delta \boldsymbol{G}^{(\ell)}$. In summary:

$$\underline{\boldsymbol{G}}^{(\ell)} \leq \boldsymbol{G}^{(\ell)} = \underline{\boldsymbol{G}}^{(\ell)} + \Delta \boldsymbol{G}^{(\ell)} \leq \overline{\boldsymbol{G}}^{(\ell)} = \underline{\boldsymbol{G}}^{(\ell)} + \overline{\Delta \boldsymbol{G}}^{(\ell)}, \tag{3.23}$$

where

$$\overline{\Delta \boldsymbol{G}}^{(\ell)} = \left(\boldsymbol{a}^{(\ell)}, ..., \boldsymbol{a}^{(\ell)}\right), \tag{3.24}$$

where $\boldsymbol{a}^{(\ell)} = \left(a_0^{(\ell)}, ..., a_p^{(\ell)}\right)^T$ is a column vector. As $\boldsymbol{G}^{(\ell)}$ is a stochastic matrix, if any element of $\overline{\boldsymbol{G}}^{(\ell)}$ is larger than 1, then we replace that element with 1.

Step 2: Bounds for $\boldsymbol{R}^{(\ell)}$, $\ell \in \mathbb{Z}^+$.

Define matrix functions (3.25a)-(3.25b) with argument matrix $\boldsymbol{X}$ for $\ell \in \mathbb{Z}^+$:

$$M_\ell(\boldsymbol{X}) \coloneqq -\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)} \boldsymbol{X}, \tag{3.25a}$$

$$N_\ell(\boldsymbol{X}) \coloneqq \boldsymbol{A}_0^{(\ell)} \boldsymbol{X}^{-1}. \tag{3.25b}$$

Comparing (3.3b) to (3.25b), one can confirm that $\boldsymbol{R}^{(\ell)} = N_\ell\left(M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)\right)$. We use these matrix functions to build bounds for $\boldsymbol{R}^{(\ell)}$ as stated below:

Proposition 3.6. $\mathbf{0} \leq \underline{\boldsymbol{R}}^{(\ell)} \leq \boldsymbol{R}^{(\ell)} \leq \overline{\boldsymbol{R}}^{(\ell)}$, where $\underline{\boldsymbol{R}}^{(\ell)} \coloneqq N_\ell\left(M_\ell\left(\underline{\boldsymbol{G}}^{(\ell+1)}\right)\right)$, $\boldsymbol{R}^{(\ell)} = N_\ell\left(M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)\right)$, and $\overline{\boldsymbol{R}}^{(\ell)} \coloneqq N_\ell\left(M_\ell\left(\overline{\boldsymbol{G}}^{(\ell+1)}\right) + \boldsymbol{D}^{(\ell)}\right)$; the non-negative diagonal matrix $\boldsymbol{D}^{(\ell)}$ is defined as:

$$\left(\boldsymbol{D}^{(\ell)}\right)_{i,j} = 0,\ i \neq j,\ \left(\boldsymbol{D}^{(\ell)}\right)_{i,i} = \begin{cases} 0 & \text{if } s_i^{(\ell)} > 0, \\ \left|s_i^{(\ell)}\right| + \xi^{(\ell)} & \text{if } s_i^{(\ell)} \leq 0, \end{cases} \tag{3.26}$$

where $|.|$ is the absolute value of a number, $s_i^{(\ell)}$ is the i-th row sum of $M_\ell\left(\overline{\boldsymbol{G}}^{(\ell+1)}\right)$ and $\xi^{(\ell)}$ is a scalar in the open range of $\left(0, \min_{i \in \mathbb{Y}}\left\{r^{(\ell)}{}_i\right\}\right)$ given that $r^{(\ell)}{}_i$ is the i-th row sum of $M_\ell(\boldsymbol{I})$.

The error matrix $\overline{\Delta \boldsymbol{R}}^{(\ell)} = \overline{\boldsymbol{R}}^{(\ell)} - \underline{\boldsymbol{R}}^{(\ell)}$ is an upper bound on $\Delta \boldsymbol{R}^{(\ell)} = \boldsymbol{R}^{(\ell)} - \underline{\boldsymbol{R}}^{(\ell)}$.

Step 3: Bounds for $\boldsymbol{x}_0$.

Let $\boldsymbol{y} = (y_0, ..., y_p)$ be a row vector of unknowns and define $\boldsymbol{Z} = \boldsymbol{A}_0^{(0)}\boldsymbol{G}^{(1)}$. Latouche and Ramaswami (1999, Theorem 12.1.4), show that $\boldsymbol{R}^{(0)}\boldsymbol{A}_2^{(1)} = \boldsymbol{A}_0^{(0)}\boldsymbol{G}^{(1)}$. Bright and Taylor (1995) show that $\boldsymbol{y}\left(\boldsymbol{A}_1^{(0)} + \boldsymbol{R}^{(0)}\boldsymbol{A}_2^{(1)}\right) = \boldsymbol{0}$, or equivalently $\boldsymbol{y}\left(\boldsymbol{A}_1^{(0)} + \boldsymbol{Z}\right) = \boldsymbol{0}$, has a positive solution $\boldsymbol{y} > \boldsymbol{0}$ such that $\boldsymbol{y}\boldsymbol{1} = 1$. We expand the system $\boldsymbol{y}\left(\boldsymbol{A}_1^{(0)} + \boldsymbol{Z}\right) = \boldsymbol{0}$ as:

$$(y_1, ..., y_{p+1}) \left(\begin{pmatrix} -a_{1,1} & \dots & a_{1,p+1} \\ \vdots & \ddots & \vdots \\ a_{p+1,1} & \dots & -a_{p+1,p+1} \end{pmatrix} + \begin{pmatrix} z_{1,1} & \dots & z_{1,p+1} \\ \vdots & \ddots & \vdots \\ z_{p+1,1} & \dots & z_{p+1,p+1} \end{pmatrix} \right) =$$

$$(0, ..., 0). \tag{3.27}$$

To solve (3.27) we fix one variable and remove one equation and then normalize the solution as described by Baumann and Sandmann (2012). If $\boldsymbol{y}$ is a solution for (3.27), then $\boldsymbol{x}_0 = \alpha\boldsymbol{y}$, $\forall\alpha \in \mathbb{R}$, is also a solution for (3.27). We choose α such that $x_{0,p+1} = 1$ and remove the last equation, and re-write (3.27) as:

$$(x_{0,1}, ..., x_{0,p}) \left(\begin{pmatrix} -a_{1,1} & \dots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{p,1} & \dots & -a_{p,p} \end{pmatrix} + \begin{pmatrix} z_{1,1} & \dots & z_{1,p} \\ \vdots & \ddots & \vdots \\ z_{p,1} & \dots & z_{p,p} \end{pmatrix} \right) =$$

$$- (a_{p+1,1} + z_{p+1,1}, ..., a_{p+1,p} + z_{p+1,p}). \tag{3.28}$$

As $\boldsymbol{G}^{(1)}$ is unknown, we cannot compute the exact value of $\boldsymbol{Z}$, and in turn can not solve (3.28). Instead, we use $\underline{\boldsymbol{Z}} = \boldsymbol{A}_0^{(0)}\underline{\boldsymbol{G}}^{(1)}$ and $\overline{\boldsymbol{Z}} = \boldsymbol{A}_0^{(0)}\overline{\boldsymbol{G}}^{(1)}$ in Proposition 3.7 to provide a lower bound and an upper bound for $\boldsymbol{x}_0$.

Proposition 3.7. $\boldsymbol{0} < \underline{\boldsymbol{x}}_0 \leq \boldsymbol{x}_0 \leq \overline{\boldsymbol{x}}_0$, If $\underline{\boldsymbol{x}}_0$ and $\overline{\boldsymbol{x}}_0$ are solutions of $\underline{\boldsymbol{x}}_0\underline{\boldsymbol{K}} = \boldsymbol{0}$ and $\overline{\boldsymbol{x}}_0\left(\overline{\boldsymbol{K}} + \boldsymbol{H}\right) = \boldsymbol{0}$, respectively, where $\underline{\boldsymbol{K}} = \boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0^{(0)}\underline{\boldsymbol{G}}^{(1)}$ and $\overline{\boldsymbol{K}} = \boldsymbol{A}_1^{(0)} +$

$\boldsymbol{A}_0^{(0)}\overline{\boldsymbol{G}}^{(1)}$; the non-negative diagonal matrix $\boldsymbol{H}$ is defined as:

$$(\boldsymbol{H})_{i,j} = 0,\ i \neq j,\ \forall i,j \in \mathbb{Y}, \tag{3.29}$$

$$(\boldsymbol{H})_{p+1,p+1} = 0,\ \underset{\forall i \in \mathbb{Y}\setminus\{p+1\}}{(\boldsymbol{H})_{i,i}} = \begin{cases} 0 & \text{if } g_i > 0, \\ (-g_i - \xi)/\underline{\boldsymbol{x}}_i & \text{if } g_i \leq 0, \end{cases} \tag{3.30}$$

where

$$g_i = \left(\underline{\boldsymbol{x}}\left(\boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0^{(0)}\overline{\boldsymbol{G}}^{(1)}\right)\right)_i, \tag{3.31}$$

and $\xi > 0$.

As a result of Proposition 3.7, $\overline{\Delta}\boldsymbol{x}_0 \coloneqq \overline{\boldsymbol{x}}_0 - \underline{\boldsymbol{x}}_0$ is an upper bound on $\Delta\boldsymbol{x}_0 \coloneqq \underline{\boldsymbol{x}}_0 - \boldsymbol{x}_0$.

Step 4: Bounds for $\boldsymbol{x}_\ell$, $\ell \in \mathbb{Z}^{++}$.

We combine $\underline{\boldsymbol{x}}_0$ and $\overline{\boldsymbol{x}}_0$ from Proposition 3.7 with recursions $\underline{\boldsymbol{x}}_{\ell+1} = \underline{\boldsymbol{x}}_\ell\ \underline{\boldsymbol{R}}^{(\ell)}$ and $\overline{\boldsymbol{x}}^{\ell+1} = \overline{\boldsymbol{x}}_\ell\ \overline{\boldsymbol{R}}^{(\ell)}$, for $\ell \in \mathbb{Z}^+$, to obtain bounds on $\boldsymbol{x}_\ell$:

Proposition 3.8. $\forall \ell \in \mathbb{Z}^{++}$, $\mathbf{0} \leq \underline{\boldsymbol{x}}_\ell \leq \boldsymbol{x}_\ell \leq \overline{\boldsymbol{x}}_\ell$.

Step 5: Bounds for $\boldsymbol{\pi}_\ell$.

For a truncation level $k \geq \tau$, we calculate bounds on the normalizing coefficient c and obtain bounds on $\boldsymbol{\pi}_\ell$. Using Proposition 3.8, we find a lower bound $\underline{c}$ for c:

$$c = \left(\sum_{\ell=0}^{k} \boldsymbol{x}_\ell\right)\mathbf{1} + \left(\sum_{\ell=k+1}^{\infty} \boldsymbol{x}_\ell\right)\mathbf{1} \geq \left(\sum_{\ell=0}^{k} \boldsymbol{x}_\ell\right)\mathbf{1} \geq \left(\sum_{\ell=0}^{k} \underline{\boldsymbol{x}}_\ell\right)\mathbf{1} =: \underline{c}. \tag{3.32}$$

To obtain an upper bound $\overline{c}$ on c, we use the rate matrix of Level τ higher than which the elements of rate matrices decrease by increasing the level; that is $\boldsymbol{R}^{(\tau+i+1)} \leq \boldsymbol{R}^{(\tau+i)}$ for $i \in \mathbb{Z}^{++}$. Finding τ is an open problem; we use a heuristic to

find it as discussed in Section 3.7. If $\ell > \tau$ then:

$$\prod_{i=1}^{n} \boldsymbol{R}^{(\ell+i)} \leq \left(\boldsymbol{R}^{(\ell+1)}\right)^n, \ n \geq 1. \tag{3.33}$$

We expand c as we did in (3.18), and use the inequalities from Propositions 3.7 and 3.8, and (3.33) as follows:

$$\begin{aligned}
c &= \left(\sum_{\ell=0}^{k} \boldsymbol{x}_\ell\right) \mathbf{1} + \left(\sum_{\ell=k+1}^{\infty} \boldsymbol{x}_\ell\right) \mathbf{1} && (3.34)\\
&= \left(\sum_{\ell=0}^{k} \boldsymbol{x}_\ell\right) \mathbf{1} + \boldsymbol{x}_{k+1} \left(I + \boldsymbol{R}^{(k+1)} + \boldsymbol{R}^{(k+1)}\boldsymbol{R}^{(k+2)} + ...\right) \mathbf{1} && (3.35)\\
&\leq \left(\sum_{\ell=0}^{k} \overline{\boldsymbol{x}}_\ell\right) \mathbf{1} + \overline{\boldsymbol{x}}_{k+1} \left(I + \boldsymbol{R}^{(k+1)} + \boldsymbol{R}^{(k+1)^2} + ...\right) \mathbf{1} && (3.36)\\
&= \left(\sum_{\ell=0}^{k} \overline{\boldsymbol{x}}_\ell\right) \mathbf{1} + \overline{\boldsymbol{x}}_{k+1} \left(I - \boldsymbol{R}^{(k+1)}\right)^{-1} \mathbf{1} && (3.37)\\
&\leq \left(\sum_{\ell=0}^{k} \overline{\boldsymbol{x}}_\ell\right) \mathbf{1} + \overline{\boldsymbol{x}}_{k+1} \left(I - \overline{\boldsymbol{R}}^{(k+1)}\right)^{-1} \mathbf{1} =: \overline{c}. && (3.38)
\end{aligned}$$

The last inequality holds if all eigenvalues of $\overline{\boldsymbol{R}}^{(k+1)}$ are strictly within the unit circle. If this condition does not hold, then we set $\overline{c} = \infty$.

Proposition 3.9. $\mathbf{0} \leq \underline{\boldsymbol{\pi}}_\ell \leq \boldsymbol{\pi}_\ell \leq \overline{\boldsymbol{\pi}}_\ell$, where $\underline{\boldsymbol{\pi}}_\ell = \underline{\boldsymbol{x}}_\ell / \overline{c}$ and $\overline{\boldsymbol{\pi}}_\ell = \overline{\boldsymbol{x}}_\ell / \underline{c}$.

3.9 Conclusion

We proposed methods to study the waiting times in an ED. We modeled the ED as an LDQBD process with infinite levels and finite phases, because it was natural to view the ED as a queueing system with multiple servers (physicians or beds) and multiple classes of customers (patients with different acuity levels) who were impatient (patients might leave the system before being seen by a physician). Our modeling approach can be applied to other health care areas like organ transplantation systems.

The methods available to compute stationary performance measures of LDQBDs

are numerical and rely on heuristically truncating the system at a level. Two main shortcomings of these methods are: 1) the truncation level is chosen heuristically, and 2) there are no error bounds on calculated performance measures.

In this study, we proposed two algorithms to address the issues associated with current solution approaches. In the first algorithm, we used Lyapunov analysis to cut the LDQBD state space such that the truncated upper tail was guaranteed to include less than a desired proportion of the probability mass. This method can be used with currently available methods to find an appropriate truncation level. In the second algorithm, we extended one of current solution methods such that the new algorithm automatically truncates the system at a level and calculates the performance measure of interest with a desired accuracy. We provided numerical examples to demonstrate our methods. However, more numerical experiments are needed to (1) compare the truncation levels that we get from the Lyapunov analysis with those of Algorithm 3.2, and (2) compare the performance of solution algorithms for LDQBDs that are available in the literature with that of Algorithm 3.2.

Another interesting direction for future research is extending our LDQBD model to systems with more than two customer types. Here, we discuss two potential approaches that one can take to extend our results to systems with n customer types, namely Class-1 (the highest priority) up to Class-n (the lowest priority) customers: (1) One can map Class-n customer counts to the level and assume that the system capacity for these customers is unlimited. One can further assume that the system capacity for all other customer classes is limited, and use all possible combinations of the number of Class-1 to Class-$(n-1)$ customers as the system phase. One potential difficulty associated with this approach is that it is not obvious how one should set the system capacity for Class-1 up to Class-$(n-1)$ customers such that desired blocking probabilities are satisfied. (2) One can use the following heuristic: Use the two-class model to analyze Class-n customers versus Class-1 to Class-(n-1) customers combined, then use the two-class model to analyze Class-n and $n-2$ customers combined versus Class-1 to Class-$(n-3)$ customers. Continue this approach until enough information is gathered.

Algorithm3.2: Computing bounds on $\boldsymbol{\pi}_{\ell'}$ for a given level $s \in \mathbb{Z}^+$.

1. Input ℓ', $\epsilon_{\ell'}$, λ_1, λ_2, μ_1, μ_2, γ_1, γ_2, c, τ, m, p, ξ,
2. Initialize $\ell = 0$, $\underline{\boldsymbol{x}}_{\ell'} = \mathbf{0}$, $\overline{\boldsymbol{x}}_{\ell'} = (\infty, ..., \infty)$,
3. Call Algorithm 3.3 to compute $\underline{\boldsymbol{R}}^{(0)}$ and $\overline{\boldsymbol{R}}^{(0)}$,
4. Solve $\underline{\boldsymbol{x}}_0 \left(\boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0^{(0)} \underline{\boldsymbol{G}}^{(1)} \right) = \mathbf{0}$ for $\underline{\boldsymbol{x}}_0$,
5. Solve $\overline{\boldsymbol{x}}_0 \left(\boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0^{(0)} \overline{\boldsymbol{G}}^{(1)} + \boldsymbol{H} \right) = \mathbf{0}$ for $\overline{\boldsymbol{x}}_0$, where
6. $(\boldsymbol{H})_{i,j} = 0, \ i \neq j, \ \forall i, j \in \mathbb{Y}$, and
7. $(\boldsymbol{H})_{p+1,p+1} = 0, \ \underset{\forall i \in \mathbb{Y} \setminus \{p+1\}}{(\boldsymbol{H})_{i,i}} = \begin{cases} 0 & \text{if } g_i > 0, \\ (-g_i - \xi)/\underline{\boldsymbol{x}}_i & \text{if } g_i \leq 0, \end{cases}$
8. $g_i = \left(\underline{\boldsymbol{x}} \left(\boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0^{(0)} \overline{\boldsymbol{G}}^{(1)} \right) \right)_i$.
9. $\underline{\boldsymbol{x}}_1 = \underline{\boldsymbol{x}}_0 \ \underline{\boldsymbol{R}}^{(0)}$ and $\overline{\boldsymbol{x}}_1 = \overline{\boldsymbol{x}}_0 \ \overline{\boldsymbol{R}}^{(0)}$,
10. Initialize normalizing factors:
11. $\underline{c} = \underline{\boldsymbol{x}}_0 \mathbf{1}$, $\overline{c}_1 = \overline{\boldsymbol{x}}_0 \mathbf{1}$,
12. If $\ell \geq \tau$ and all eigenvalues of $\overline{\boldsymbol{R}}^{(1)}$ are strictly within the unit
13. circle Then $\overline{c}_2 = \overline{\boldsymbol{x}}_1 \left(I - \overline{\boldsymbol{R}}^{(1)} \right)^{-1} \mathbf{1}$, otherwise $\overline{c}_2 = \infty$,
14. $\overline{c} = \overline{c}_1 + \overline{c}_2$,
15. $\underline{\boldsymbol{\pi}}_{\ell'} = \underline{\boldsymbol{x}}_{\ell'} / \overline{c}$ and $\overline{\boldsymbol{\pi}}_{\ell'} = \overline{\boldsymbol{x}}_{\ell'} / \underline{c}$,
16. While $(\overline{\boldsymbol{\pi}}_{\ell'} - \underline{\boldsymbol{\pi}}_{\ell'}) \mathbf{1} > \epsilon_{\ell'}$
17. $\ell = \ell + 1$,
18. Call Algorithm 3.3 to compute calculate $\underline{\boldsymbol{R}}^{(\ell+1)}$ and $\overline{\boldsymbol{R}}^{(\ell+1)}$,
19. $\underline{\boldsymbol{x}}_{\ell+1} = \underline{\boldsymbol{x}}_{\ell} \ \underline{\boldsymbol{R}}^{(\ell)}$ and $\overline{\boldsymbol{x}}_{\ell+1} = \overline{\boldsymbol{x}}_{\ell} \ \overline{\boldsymbol{R}}^{(k)}$,
20. Update normalizing factors:
21. $\underline{c} = \underline{c} + \underline{\boldsymbol{x}}_k \mathbf{1}$, $\overline{c}_1 = \overline{c}_1 + \overline{\boldsymbol{x}}_{\ell} \mathbf{1}$,
22. If $\ell \geq \tau$ and all eigenvalues of $\overline{\boldsymbol{R}}^{(\ell+1)}$ are strictly within the unit
23. circle Then $\overline{c}_2 = \overline{\boldsymbol{x}}_{\ell+1} \left(I - \overline{\boldsymbol{R}}^{(\ell+1)} \right)^{-1} \mathbf{1}$, otherwise $\overline{c}_2 = \infty$,
24. $\overline{c} = \overline{c}_1 + \overline{c}_2$,
25. $\underline{\boldsymbol{\pi}}_{\ell'} = \underline{\boldsymbol{x}}_{\ell'} / \overline{c}$ and $\overline{\boldsymbol{\pi}}_{\ell'} = \overline{\boldsymbol{x}}_{\ell'} / \underline{c}$,
26. Return
27. Output truncation level $k = \ell, \overline{\boldsymbol{\pi}}_{\ell'}, \underline{\boldsymbol{\pi}}_{\ell'}$

Algorithm3.3: Computing bounds on $\boldsymbol{R}^{(\ell)}$ for a given level $\ell \in \mathbb{Z}^+$.

1. Input: ℓ, λ_1, λ_2, μ_1, μ_2, γ_1, γ_2, c, m, p,
2. $\boldsymbol{G}_0^{(\ell+m+1)} = \boldsymbol{0}$,
3. For $i = m, ..., 1$, compute $\boldsymbol{G}_{m-i+1}^{(\ell+i)} = \left(-\boldsymbol{A}_1^{(\ell+i+1)} - \boldsymbol{A}_0^{(\ell+i+1)}\boldsymbol{G}_{m-i}^{(\ell+i+1)}\right)^{-1} \boldsymbol{A}_2^{(\ell+1)}$,
4. $\underline{\boldsymbol{G}}^{(\ell+1)} = \boldsymbol{G}_m^{(\ell+1)}$,
5. $\overline{\Delta \boldsymbol{G}}^{(\ell+1)} = \left(\boldsymbol{a}^{(\ell+1)}, ..., \boldsymbol{a}^{(\ell+1)}\right)$, where $a_i{}^{(\ell+1)} = 1 - \sum_{j=1}^{p} \left(\underline{\boldsymbol{G}}^{(\ell+1)}\right)_{i,j}$, $\forall i \in \mathbb{Y}$,
6. $\overline{\boldsymbol{G}}^{(\ell+1)} = \underline{\boldsymbol{G}}^{(\ell+1)} + \overline{\Delta \boldsymbol{G}}^{(\ell+1)}$,
7. $\underline{\boldsymbol{R}}^{(\ell)} = \boldsymbol{A}_0^{(\ell)} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\underline{\boldsymbol{G}}^{(\ell+1)}\right)^{-1}$
8. $\overline{\boldsymbol{R}}^{(\ell)} = \boldsymbol{A}_0^{(\ell)} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\overline{\boldsymbol{G}}^{(\ell+1)} + \boldsymbol{D}^{(\ell)}\right)^{-1}$, where
9. $\left(\boldsymbol{D}^{(\ell)}\right)_{i,j} = 0,\ i \neq j$, and
10. $\left(\boldsymbol{D}^{(\ell)}\right)_{i,i} = \begin{cases} 0 & \text{if } s_i^{(\ell)} > 0, \\ -s_i^{(\ell)} + \xi^{(\ell)} & \text{if } s_i^{(\ell)} \leq 0, \end{cases}$
11. $s_i^{(\ell)} = \sum_{j \in \mathbb{Y}} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\overline{\boldsymbol{G}}^{(\ell+1)}\right)_{i,j}$.
12. $\xi^{(\ell)} = \min\left\{\sum_{j \in \mathbb{Y}} \left(-\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}\right)_{i,j} \middle| i \in \mathbb{Y}\right\}/2$.

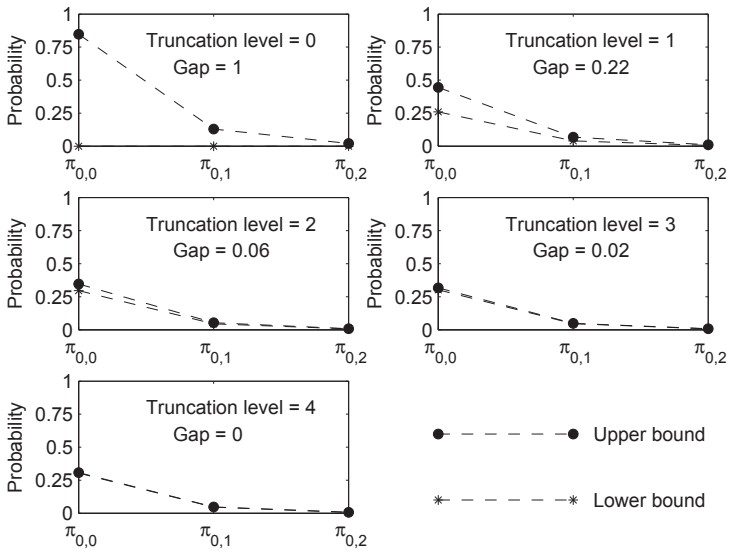


Figure 3.3: Bounds on $\boldsymbol{\pi}_0$ elements for System 4 in Table 3.5.

CHAPTER 4

Predicting the Spatial Distribution of Demand for Percutaneous Coronary Intervention in Alberta

4.1 Introduction

Accurate estimates of demand for health care services are essential in health care planning. These estimates are used for both short-term planning, surgical scheduling (Chow et al. 2011), for example, and for long-term planning, location (Gu et al. 2010) and capacity planning (Patrick et al. 2015), for example. In this chapter, we specifically focus on providing a framework for predicting the number of heart attack patients in geographical areas that are sufficiently small to provide useful estimates of travel times. This study has been inspired by the health care authorities' need for accurate demand estimates for heart attack treatment centers in Alberta.

When our heart pumps properly, it delivers oxygen and blood to every part of our body. At the same time, coronary arteries supply blood and oxygen to the heart itself. Myocardial infarctions, or heart attacks, occur when a coronary artery is partially or completely blocked by a blood clot. Once blocked, an artery no longer properly supplies oxygen to a certain part of the heart muscle, and this lack of oxygen causes a heart attack. If a clot blocks an artery completely, all parts of the heart supplied by the artery start to die, and a severe heart attack, ST-Segment Elevation Myocardial Infarction (STEMI), occurs (Fogoros 2008b). Partial blocking of an artery results in a less severe heart attack, referred to as NSTEMI (Fogoros

2008a). We use the terms "heart attack" and "STEMI" interchangeably in this research. Heart attack and other diseases and injuries of the cardiovascular system are called cardiovascular diseases (Heart & Stroke Foundation 2013). Percutaneous Coronary Intervention (PCI), restoration of the blood flow in a blocked heart artery by inserting a tube into the artery, is an effective heart attack treatment method if administered in a timely fashion (Kutcher et al. 2009, Armstrong et al. 2003).

According to the Heart & Stroke Foundation (2013), almost 70,000 heart attacks occur in Canada each year and around 16,000 of these patients die, mainly out of hospitals. During 2008, almost 6% of all deaths in Canada were attributed to heart attacks. Bakal et al. (2011) estimate based on $2002 - 2007$ Alberta Health and Wellness data that the annual rate of heart attack in Alberta is 0.8 per 1000 people. There are currently three PCI facilities in Alberta—two in Edmonton and one in Calgary. Patel et al. (2007) estimated that 70% of Albertans above the age of 20 can reach one of the current PCI facilities within the 90-minute window by ground ambulance. Therefore, almost a third of adult Albertans do not have timely access to a PCI facility, which has prompted research on adding new facilities and on more efficient usage of the current facilities. The sensitivity of this treatment method to the time that has passed after the onset of a heart attack until the operation starts, motivates health care authorities to find the best set of locations to open new PCI centers to maximize timely coverage of the population. This research provides accurate demand estimates for PCI centers as a foundation for future studies on finding the optimal locations.

In order to mitigate issues associated with data aggregation errors in future studies on finding the best locations for PCI facilities, we use the Dissemination Area (DA) as our spatial unit. DAs are small geographical areas with population sizes from 400 to 700 that cover all the territory of Canada. DAs are defined by Census Canada and are the smallest geographical units for which all census data are published (Statistics Canada 2010).

A good deal of research has been done on identifying and studying heart attack risk factors such as smoking, inactivity, high blood pressure and total cholesterol

(e.g. Wielgosz et al. 2009, Wilson et al. 1998), and the odds of having a heart attack for a specific person based on the level of her unhealthy behaviour and physical conditions (e.g. Wilson et al. 1998). Nevertheless, we can not use these heart attack risk factors as predictors of the number of heart attacks in a DA because direct information about individual-level heart attack risk factors is not readily available for DAs. For example, it is difficult to obtain information about the proportion of people with high blood pressure in a DA as opposed to in the whole province.

We use the population size in each of several cohorts, defined by age, gender, income, and education as explanatory variables which are published by Census Canada for DAs. As we discuss in Section 4.2, medical researchers have done a great deal of research on the impact of the variables that we use in this study either directly on an individual's chance of developing a hard attack or on heart attack risk factors. In contrast, we study the impact of these variables on the DA level; we use statistical models to predict the number of heart attack patients in DAs as a function of the explanatory variables.

We use Poisson regression to build predictive models of heart attack counts. Poisson regression assumes that the dependent variable has a mean that is a function (the "link function") of the explanatory variables. We show that the commonly used exponential link function has undesirable properties in our setting and instead propose using the identity link function, together with constraints on parameter values to ensure that the predicted Poisson means are positive. We compare the predictive power of Poisson regression with that of a standard multiple linear regression approach.

4.2 Literature Review

Heart diseases are significantly more common in middle-aged men than in women. Heart attack rates increase for both men and women with age but the increase is steeper for women (Hulley et al. 1998, Jousilahti et al. 1999, Martins et al. 2001, Albert et al. 2006).

Other than age and sex, there are other risk factors for cardiovascular diseases: Blood pressure; smoking; total, LDL, and HDL cholesterol (Gordon and Kannel 1982); family history; and obesity (Grundy et al. 1993). Information about these risk factors is not readily available at the DA level.

Lack of education and income are associated with higher risk factor levels for cardiovascular diseases (Winkleby et al. 1992, Diez-Roux et al. 2000). The correlation between education level, income and cardiovascular risk factors are stronger in high-income countries like Canada (Rosengren et al. 2009, Goyal et al. 2010). Information about the education and income levels of the population are available at the DA level in Canada.

4.3 Data

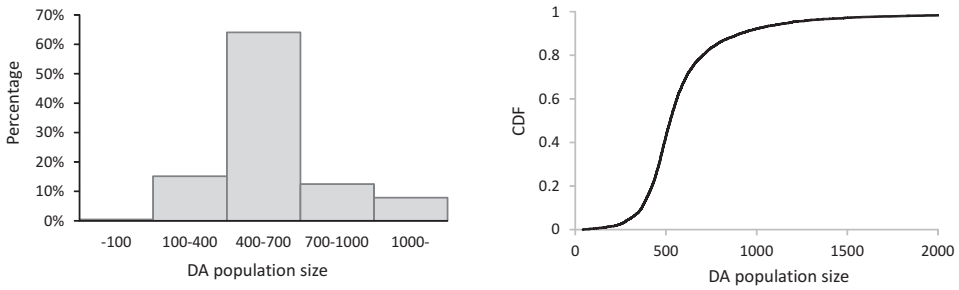
Statistics Canada publishes census data every 5 years—the last one was in 2011. We use the 2006 census (as opposed to 2011) to obtain information about age, sex, education, and income variables for each DA in Alberta because 2006 falls within the 2003-2010 period for which we have heart attack incidence data, as discussed below. In the 2006 census, Alberta was covered by 5209 DAs. As shown in Figure 4.1, the population size of almost 40% of the DAs did not fall into the 400-700 range, the population size of a DA as defined by Statistics Canada (2010). This could be because, for each census, the population count from the previous census is used to define DA borders, so the process of defining DA borders is always lagging behind the process of counting actual population of the DAs. Descriptive statistics of the population of DAs is presented in Table 4.1. As we discuss later in Section 4.4, the high variation in the DA population adds to the complexity of our prediction process.

The 2006 census provides information about the number of people in the following cohorts:

- Males, by 5-year age groups (0-5, 6-10, ..., 76-80, and over 85),
- Females, by 5-year age groups (0-5, 6-10, ..., 76-80, and over 85),

Table 4.1: Alberta DAs' population and heart attack incidence per DA.

Statistics	DAs	STEMIs per DA
Total number	5,209	14,287
Mean	628.48	2.74
Standard Deviation	535.30	4.27
Median	524	2
Maximum	11,881	74



(a) Population distribution histogram.

(b) Population distribution CDF.

Figure 4.1: The population distribution in Alberta DAs in 2006 census.

- Males with post-secondary degree,
- Females with post-secondary degree,
- Those with low income that are over 65. Statistics Canada (2010) defines a low-income person or family as one "who spend 20% more than average of their before-tax income on food, shelter and clothing."

We build and validate our models using empirical heart attack data from Alberta. Our data set includes 14,287 heart attack incidents that occurred in Alberta from 2003-2010 by postal code (PC). A PC consists of one or more postal addresses that Canada Post has assigned a single six-character alphanumeric code, for example, T6G 2R6 (Canada Post 2012). We aggregate the PC-level heart attack incidence data to the DA level because our explanatory variables are at the DA level. Figure 4.2 shows a heat map of the rate of STEMI patients in DAs across the province. Total of 14,287 STEMIs over 8 years spread over 5,209 DAs with average population of 628.48 translates into an average of approximately one heart attack per 1,833 people per year.

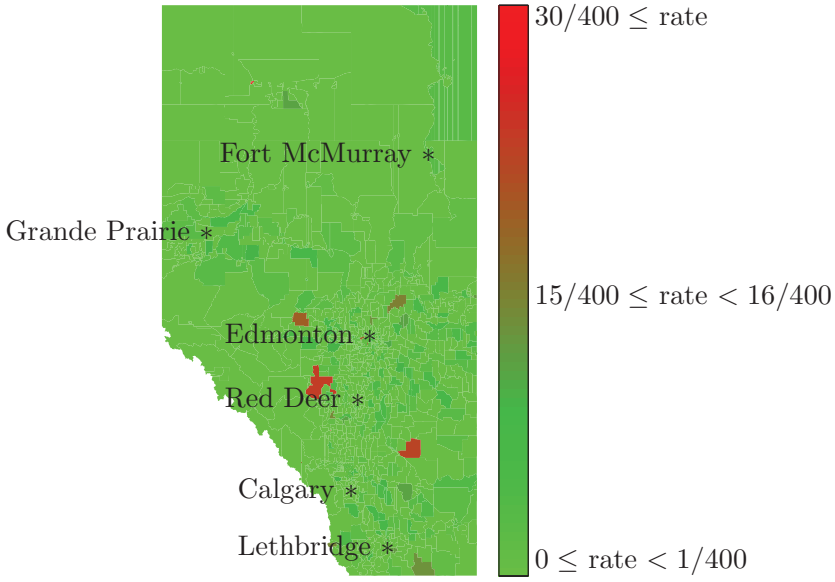


Figure 4.2: Heat map for STEMI incidents in Alberta DAs.

Figure 4.3 shows histogram of the proportion of DAs with different number of observed STEMIs in Alberta over the study period. Note that more than 30% of the DAs did not have any STEMI patients from 2003 to 2010. Figure 4.4 shows the scatter plot of STEMI counts versus DA populations with a trend line that has been estimated by no-intercept simple linear regression. The slope of this trend line, 0.0038, is the predicted chance of an individual in Alberta to experience a heart attack over an 8-year period given that the total population is the only explanatory variable. If we divide this number by the duration of study period, $0.0038/8 = 0.00047$, we then obtain a rate close to what we roughly calculated in the previous paragraph, $1/1{,}833 = 0.00054$.

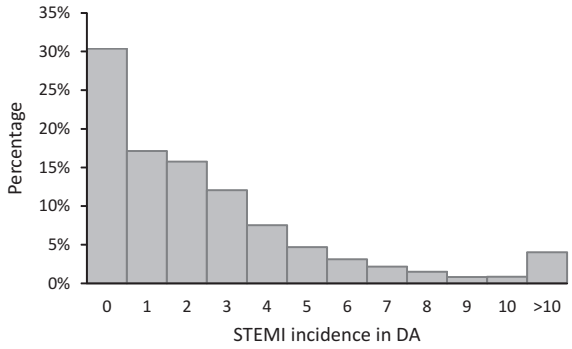


Figure 4.3: STEMI distribution in Alberta DAs.

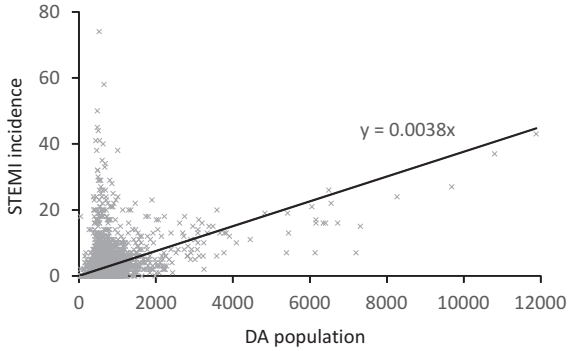


Figure 4.4: STEMI distribution in Alberta DAs.

4.4 Model Specification

If X is the total population of a DA, and this is the only information available, then a natural formula to predict the number Y of STEMI patients in that DA is $Y = bX$, where b is the probability that an individual will experience a STEMI incident. If we know the number X_1 of females and number X_2 of males in that DA, then $Y = b_1X_1 + b_2X_2$ is a natural formula to predict Y, with b_1 and b_2, the probabilities that a female and a male will experience a STEMI incident, respectively. Note that these two equations have the desirable properties that (1) A DA with zero population has zero predicted STEMI incidents and (2) if the DA population increases by a given percentage, holding the gender ratio constant, the predicted number of STEMI incidents increases by the same percentage. In other words, these equations are invariant to scale.

In general, if one has information about the population of each Cohort j, where $j = 1, ..., n$, in each DA, then one can estimate the number of STEMI patients with a similar function, $Y = \sum_{i=1}^{n} b_i X_i$, using standard multiple linear regression with the intercept forced to 0. However, Y is integer-valued and often 0, and therefore we also investigate the use of count regression models such as Poisson regression and binomial regression.

In our models, each DA is an observation where the number Y_i of STEMI patients in DA i, $i \in \{1, \ldots, m\}$ is the dependent variable and the number X_{ij} of people in Cohort j, for $j \in \{1, ..., n\}$, of DA i are the explanatory variables. Let X_i be the population of DA i, $i \in \{1, \ldots, m\}$. Ideally, the cohorts would be mutually exclusive and collectively exhaustive, which would imply $X_i = \sum_{j=1}^{n} X_{ij}$, for all $i \in \{1, \ldots, m\}$, but unfortunately it is not always possible to define the cohorts to satisfy these conditions because of data limitations.

To predict the number of STEMI incidents in a DA using Poisson regression, we assume that Y_i, $i \in \{1, \ldots, m\}$, has a Poisson distribution with mean λ_i, which is related to a linear combination of the explanatory variables X_{ij}, $j \in \{1, \ldots, n\}$, through a function $f(.)$. The function $f(.)$ is referred to as the link function. It follows that, for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, the probability of observing y_i STEMI incidents in DA i is:

$$\Pr(Y_i = y_i | X_{ij} = x_{ij}) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \ \lambda_i = f\left(b_0 + \sum_{j=1}^{n} b_j x_{ij}\right). \tag{4.1}$$

We will use the "hat" symbol to denote the estimated parameters.

The most common link function is the exponential function (Greene 2012). That is, $\lambda_i = \exp(b_0 + \sum_{j=1}^{n} b_j x_{ij})$. King (1988) argues that the exponential link function is usually appropriate, at least for applications in political science, because it guarantees that the predicted Poisson means are positive and are well interpreted. However, we show that the exponential link function has undesirable properties when used for the count of STEMI patients in a DA. We demonstrate those undesirable properties by focusing on $\mathrm{E}[Y]/X = \lambda/X$, the mean number of heart attacks

per person for a DA with population X.

Let $p_{ij} = X_{ij}/X_i$ be the proportion of the population in DA i that belongs to Cohort j. We postulate that the following is a desirable property for an equation relating the X_{ij} to Y_i:

Assumption 4.1. *If p_{ij}, for all $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, do not change when the population X_i changes, then the per-capita heart attack rate $E[Y_i]/X_i = \lambda_i/X_i$ is independent of X_i, that is,*

$$\frac{\partial\left(\frac{\lambda_i}{X_i}\right)}{\partial X_i} = 0, \ \forall i \in \{1, ..., m\}. \tag{4.2}$$

According to Assumption 4.1, if DAs 1 and 2 have identical cohort proportions, that is, $p_{1j} = p_{2j}$, for all $j \in \{1, \ldots, n\}$, then the two DAs should have identical predicted heart attack rates. We use the following two examples to show that the exponential link function does not satisfy Assumption 4.1.

Example 4.1. Returning to the example where the only explanatory variable is the total population, X_i, if DAs 1 and 2 both have total population x, then all individuals in both DAs have the same predicted chance of experiencing a heart attack: $\lambda/x = \exp(b_0 + b_1 x)/x$. If one merges DAs 1 and 2 to form a new DA with total population $2x$, then the predicted chance of experiencing a heart attack for an individual in the new DA is $\lambda/(2x) = \exp(b_0 + 2b_1 x)/x$, which is $\exp(b_1 x)/2$ times larger than what it was before merging the DAs.

Example 4.2. Extending Example 4.1 to the case where there are two explanatory variables, namely the number X_{i1} of females and the number X_{i2} of males, if DAs 1 and 2 both have the same population x_1 of females and population x_2 of males, then the predicted chance of heart attack for an individual in each of these DAs is $\lambda/x = \exp(b_0 + b_1 x_1 + b_2 x_2)$ before merging the DAs, and this chance is $\lambda/(2x) = \exp(b_0 + 2b_1 x_1 + 2b_2 x_2)$ after merging the DAs. This is $\exp(b_1 x_1 + b_2 x_2)/2$ times larger than what it was before merging the DAs.

As demonstrated by Examples 4.1-4.2, if we use the exponential link function,

then having fixed cohort proportions does not guarantee that the predicted per-capita heart attack rates are independent of the population size. In the count regression literature, scholars usually use *exposure* (Greene 2012) to address this issue by assuming that the rate at which counts occur, as opposed to the counts, are a function of the explanatory variables. That is, instead of $\lambda_i = f\left(b_0 + \sum_{j=1}^{n} b_j x_{ij}\right)$, we assume $\lambda_i / x_i = f\left(b_0 + \sum_{j=1}^{n} b_j x_{ij}\right)$ in (4.1), where x_i is referred to as the exposure of DA i. Therefore, if the link function is exponential, then:

$$\lambda_i = x_i f\left(b_0 + \sum_{j=1}^{n} b_j x_{ij}\right) = x_i \exp\left(b_0 + \sum_{j=1}^{n} b_j x_{ij}\right) = \exp\left(b_0 + \sum_{j=1}^{n} b_j x_{ij} + \ln x_i\right). \tag{4.3}$$

We see that if the link function is exponential, then the exposure approach corresponds to adding $\ln(X_i)$ as an explanatory variable, with a coefficient that is forced to equal 1. Poisson regression with population exposure does not satisfy Assumption 4.1, except in a special case where the model includes none of the cohort population explanatory variables that we have defined. To demonstrate this statement, we use a new example and continuations of Examples 4.1-4.2.

Example 4.3. If there are no cohort population explanatory variables, then (4.3) reduces to $\lambda_i = \exp(b_0) x_i$, which is in the form of the function $Y = bX$ that we obtained at the beginning of this section. The predicted heart attack rate of an individual is fixed and equals $\exp(b_0)$.

Example 4.4. *Example 4.1 continued.* If we apply the Poisson regression with exposure, then the before-merging and after-merging predicted probability of heart attack for individuals in both DAs 1 and 2 respectively are $\lambda / X = \exp\left(b_0 + b_1 X\right)$ and $\lambda / (2X) = \exp\left(b_0 + 2b_1 X\right)$, with the latter being $\exp\left(b_1 X\right)$ times larger than the former.

Example 4.5. *Example 4.2 continued.* Following the same fashion as in Example 4.4, one can confirm that if there are two variables and one applies the Poisson regression with exposure, then the predicted heart attack rate for individuals after

merging DAs 1 and 2 will be $\exp(b_1X_1 + b_2X_2)$ times larger than their before-merging predicted heart attack chance.

We propose an identity link function ($f(x) = x$) with a linear, without intercept, combination of the explanatory variables, that is,

$$\lambda_i = \sum_{j=1}^{n} b_j X_{ij}, \ \forall i \in \{1, ..., m\}. \tag{4.4}$$

If we use (4.4) as the link function for Examples 4.1 and 4.2, then individuals in both DAs 1 and 2 will have predicted heart attack chance of $\lambda/X = (b_1X)/X = b_1$ and $\lambda/X = (b_1X_1 + b_1X_2)/X = b_1p_1 + b_2p_2$, respectively, which do not vary when the DAs are merged.

Theorem 4.1. *The only model specification that satisfies Assumption 4.1 consists of an identity link function of a linear, without intercept, combination of the explanatory variables.*

Proof. It follows from Assumption 4.1 that, for all $i \in \{1, ..., m\}$:

$$\frac{\partial(\frac{\lambda_i}{X_i})}{\partial X_i} = \frac{\partial\left(\frac{f(X_{i1},...,X_{in})}{X_i}\right)}{\partial X_i} = \frac{\frac{\partial f(X_{i1},...,X_{in})}{\partial X_i} X_i - f(X_{i1}, \ldots, X_{in})}{X_i^2} = 0. \tag{4.5}$$

Setting the numerator of the third ratio in (4.5) equal to zero, we obtain:

$$\frac{\partial f(X_{i1}, \ldots, X_{in})}{f(X_{i1}, \ldots, X_{in})} = \frac{\partial X_i}{X_i}. \tag{4.6}$$

We take integrals of both sides of (4.6):

$$\ln(f(X_{i1}, \ldots, X_{in})) = \ln(X_i) + c_i, \tag{4.7}$$

where $c_i = \ln(f(p_{i1}, \ldots, p_{in}))$, which is obtained by setting $X_i = 1$, which implies $X_{ij} = p_{ij}$, in (4.7). It follows that:

$$f(X_{i1}, \ldots, X_{in}) = f(p_{i1}X_i, \ldots, p_{in}X_i) = f(p_{i1}, \ldots, p_{in})X_i, \tag{4.8}$$

where the first equality results from the definition of p_{ij} and the second one is resulted by replacing the value of c_i in (4.7).

We proceed by showing that $\partial f(X_{i1}, \ldots, X_{in})/\partial X_{ij}$ is constant for all $j \in \{1, ..., n\}$. If one multiplies the right-hand side of (4.6) with $\partial X_{ij}/\partial X_{ij}$, then one obtains:

$$\frac{\partial f(X_{i1}, \ldots, X_{in})}{\partial X_{ij}} = \frac{f(X_{i1}, \ldots, X_{in})}{X_i} \frac{\partial X_i}{\partial X_{ij}} = \frac{f(p_{i1}, \ldots, p_{in})}{p_{ij}}. \tag{4.9}$$

The last equality in (4.9) results from the definition of p_{ij} and (4.8).

Therefore, if the function $f(.)$ satisfies Assumption 4.1, then it follows that $f(.)$ also satisfies (4.8) and (4.9), which means that $f(.)$ is an identity link function of a linear, without intercept, combination of the explanatory variables. □

To estimate the parameter's of the Poisson regression where the expected value of its dependent variable is calculated by (4.4), we use the maximum likelihood estimation (MLE) method. Our likelihood function is:

$$\mathcal{L}(y_1, \ldots, y_m | b_1, \ldots, b_n) = \prod_{i=1}^{m} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{m} \frac{e^{-(\sum_{j=1}^{n} b_j x_{ij})} (\sum_{j=1}^{n} b_j x_{ij})^{y_i}}{y_i!}. \tag{4.10}$$

The MLE estimates, $\hat{b}_1, \ldots, \hat{b}_n$, of model parameters is the set of parameters that maximizes (4.10). As Marschner (2010) discusses, standard methods for calculating the MLE estimates that maximize (4.10) are numerically unstable as (4.4) may produce negative Poisson means. Because of the instability issue, we cannot use standard statistical software like R to compute the MLE estimates.

To compute the MLE estimates, we first derive the log-likelihood function:

$$\begin{aligned} \ln(\mathcal{L}(y_1, \ldots, y_m | b_1, \ldots, b_n)) &= \sum_{i=1}^{m} \ln\left(\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}\right) = \sum_{i=1}^{m} (y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)) \\ &= \sum_{i=1}^{m} \left(y_i \ln\left(\sum_{j=1}^{n} b_j x_{ij}\right) - \left(\sum_{j=1}^{n} b_j x_{ij}\right) - \ln(y_i!) \right). \end{aligned} \tag{4.11}$$

We proceed by building a non-linear program that maximizes (4.11) and satisfies the $\lambda_i \geq 0$ constraint, for all $i \in \{1, ..., m\}$:

$$\begin{aligned} &\text{maximize } \ln(\mathcal{L}(y_1, \ldots, y_m | b_1, \ldots, b_n)) \\ &\text{subject to } \lambda_i = \sum_{j=1}^{n} b_j X_{ij} \geq 0, \quad \forall i \in \{1, ..., m\}. \end{aligned} \tag{4.12}$$

The objective function is a strictly concave function of the decision variables $b_1, \ldots, b_n$, because the first sum in (4.11) has three components: (1) a logarithmic expression, which is a strictly concave function of the decision variables, (2) a linear expression, and (3) a constant expression, and the sum of concave, linear, and constant functions is concave. The maximization problem (4.12) has a concave objective function with linear constraints, and therefore, a local optimum solution for (4.12) is a global optimum solution (Bazaraa et al. 2006).

We use the Knitro (2015) add-on in Matlab to solve (4.12) and measure the prediction power of the Poisson regression using the Akaike Information Criterion (AIC) statistic: AIC $= 2n - 2\ln(\mathcal{L})$, where $\ln(\mathcal{L})$ is the maximized value of the log-likelihood function. Shmueli (2010) and others argue that AIC is an appropriate measure to compare predictive statistical models. Models with lower AIC values are considered better, corresponding to either better goodness of fit (higher likelihood) or a smaller number of parameters, or both.

We compare the prediction power of the Poisson regression that we developed with standard multiple linear and negative binomial regressions. The binomial regression is an extension of the Poisson regression where the Poisson mean is a function of a linear combination of the explanatory variables and a random residual which has the gamma distribution. That is,

$$\lambda_i = \sum_{j=1}^{n} b_j x_{ij} + \epsilon_i, \ i \in \{1, ..., m\}, \tag{4.13}$$

where the error term ϵ follows the gamma distribution (Greene 2012). In the same fashion as we did for Poisson regression, we build a non-linear program for maxi-

mizing the negative binomial log-likelihood function with constraints that guarantee positive means. However, the log-likelihood function for negative binomial regression is more complicated than that of Poisson regression, and we do not know whether it is concave.

4.5 Results

We have $m = 5209$ DAs, and for each DA, we have the count of 0-85 year-old males and females in 5-year groups 1-5,..., 81-85, and also the number of over-85 males and females. The number of males and females with post-secondary education and the number of low-income people are also available for each DA. In total, we have 39 variables for each DA. To find a good combination of these variables that gives a good prediction power to our Poisson model, we use a variation of the full enumeration approach as explained below. We build a large number of models and use AIC to compare their prediction power.

If we were to use the full enumeration approach, we would have to compare the AICs of more than 10 billion models. To calculate the lower bound, 10^{10}, for the total number of models, assume that we only have age-gender variables. If we combine male and female age variables to unisex variables, then we will have eighteen 5-year age group variables (1-5,...,81-85, over 85). We use cut-points to count the number of models we can build using these variables: If there are no cut-points, there is one model; if there is one cut-point, there are 17 models; if there are two cut-points, there are $17 \times 16/2$ models, etc. Therefore, there are $\sum_{k=0}^{17} \binom{17}{k} = 131,072$ models. If we add gender to our models, then we will have 131,072 options for male variables and 131,072 options for female variables, or a total of $131,072 \times 131,072 = 17,179,869,184 \approx 1.7 \times 10^{10}$ models. The total number of possible models becomes much larger than 10^{10} if one also considers the education and income variables.

To limit the number of models that we build and compare, we reduce the number of age categories to 11 by combining male and female 1-5,..., 36-40 age groups into

male 1-40 and female 1-40 age groups, respectively. We pool all under-40 age groups into one variable because people in this age range are much less likely to get a heart attack than those who are over 40. We also assume that males and females have the same age groups. Therefore, the total number of models with only age variables; age and education variables; age and income variables; age, education, and income variables, where each category can have two possibilities of with or without gender, reduces to $1024 \times 8 = 8192$.

Among those models that we constructed, a model with 15 variables has the minimum AIC, 27,328. Variables of this model are the numbers of people in the following cohorts: 0-45, 46-50, 51-55, 56-60, 61-70, and over 70, separately for males and females; males and females with post-secondary education; and those with low income. Table 4.2 shows the coefficients of the explanatory variables when we use Poisson regression with the identity link function of a linear combination of the explanatory variables, with no intercept. To evaluate the sensitivity of estimated parameters to the data, we randomly divide our DAs into 10 groups of almost the same size without replication and fit our model to each of these 10 data sets. Figure 4.5 shows that the estimated parameters are stable across different subsets of the data.

To evaluate the prediction power of the Poisson regression, we compare it's AIC with that of standard multiple linear and negative binomial regressions, when the explanatory variables are as described in Table 4.2—we refer to these models as "reduced" models. We also compare the three regression methods for a model that includes all 25 possible variables—we refer to these models as "full" models. The results are shown in Table 4.3.

Comparing AICs in Tables 4.2 and 4.3, one can confirm that the AICs of full models are slightly larger than that of the reduced models, but there are big differences across AICs of different regression methods. Although negative binomial has the minimum AIC for both reduced and full set of variables, Poisson regression is more favorable because we do not know about the concavity of the negative binomial log-likelihood function, and the mathematical program that we use to find the best

Table 4.2: Comparing regression models using a selected set of variables.

Variable	Poisson	Linear	Binomial*
Male, 0-45	0.0077	0.0075	0.0072
Male, 46-50	-0.0099	-0.0071	-0.0097
Male, 51-55	0.0021	0.0001	0.0036
Male, 56-60	-0.0098	0.0020	-0.0147
Male, 61-70	0.0105	0.0186	0.0061
Male, 71-	0.0164	0.0197	0.0121
Female, 0-45	-0.0002	0.0011	-0.0008
Female, 46-50	0.0078	0.0030	0.0091
Female, 51-55	0.0041	-0.0005	0.0068
Female, 56-60	0.0341	0.0187	0.0413
Female, 61-70	0.0084	0.0103	0.0078
Female, 71-	0.0299	0.0253	0.0338
Male, educated	-0.0009	-0.0017	-0.0005
Female, educated	-0.0075	-0.0078	-0.0067
Low income	-0.0102	-0.0127	-0.0085
AIC	27,328	28,635	21,381

* It is the best found solution.

set of coefficients for negative binomial (1) do not return the optimal solution and (2) its final solution highly depends on the initial solution provided by the user; we consistently got the best results when we fed the optimal solution of Poisson regression model into the negative binomial model as an initial solution.

Figures 4.6 (a)-(c) demonstrate that the coefficients of Poisson, linear, and negative binomial regressions are similar across the reduced and full models. In particular, the coefficients of the aggregated age categories in the reduced model are similar to the un-aggregated coefficients in the full model.

Looking at Figures 4.6 (a)-(c), we observe two counter-intuitive facts: (1) for some age groups, the coefficients are negative, meaning that if we add people to these cohorts in a DA, then we will reduce the predicted number of heart attacks in that DA, which is counter-intuitive because each person has a positive probability of having a heart attack, and (2) the coefficients of the age variables do not monotonically increase by age. Explaining these patterns is an important topic for future work.

We close with an empirical illustration of the importance of not using an expo-

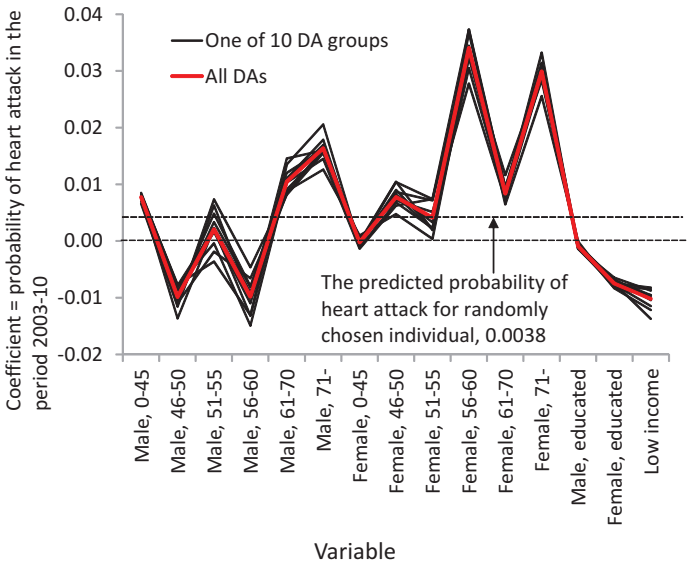


Figure 4.5: The sensitivity of model parameters to different data sets.

nential link function. The most populous DA in Alberta is located in Red Deer with a population size of almost 11,890 people and 43 heart attacks during the study period—remember that, as discussed in Section 4.3, the population sizes of some DAs do not agree with the targeted 400-700 people for DAs, possibly because of rapid population growth. Figure 4.7 shows the location of Red Deer in the province and a close up of the most populous DA. When we used the exponential link function, our prediction for the number of STEMI incidents became 210.4 (more than 4 times the real number, 43). When we split the most populous DA into 20 fictional homogeneous DAs with equal population size of $11,890/20 = 594.5$, the prediction for each fictional DA became 1.31. Now the prediction of the total number of heart attacks, $20 \times 1.31 = 26.3$ appears more reasonable. This shows how prediction using the exponential link function depends on the spatial units defined by the government. When we used the identity link function and no intercept, the prediction for the whole DA became 39.18 heart attacks.

Table 4.3: Comparing regression models using all variables.

	Poisson	Linear	Binomial*
Male, 0-40	0.0075	0.0077	0.0071
Male, 41-45	0.0084	0.0049	0.0072
Male, 46-50	-0.0089	-0.0069	-0.0076
Male, 51-55	0.0022	0.0003	0.0032
Male, 56-60	-0.0097	0.0016	-0.0144
Male, 61-65	0.0068	0.0139	0.0027
Male, 66-70	0.0166	0.0265	0.0126
Male, 71-75	0.0129	0.0290	-0.0002
Male, 76-80	0.0136	0.0208	0.0113
Male, 81-85	0.0163	0.0119	0.0184
Male, 86-	0.0244	0.0167	0.0256
Female, 0-40	0.0000	0.0011	-0.0005
Female, 41-45	-0.0035	0.0015	-0.0051
Female, 46-50	0.0086	0.0038	0.0100
Female, 51-55	0.0044	0.0001	0.0069
Female, 56-60	0.0350	0.0199	0.0420
Female, 61-65	0.0045	0.0047	0.0064
Female, 66-70	0.0152	0.0135	0.0139
Female, 71-75	0.0284	0.0140	0.0369
Female, 76-80	0.0224	0.0205	0.0248
Female, 81-85	0.0384	0.0435	0.0385
Female, 86-	0.0265	0.0205	0.0300
Male, educated	-0.0008	-0.0016	-0.0003
Female, Educated	-0.0075	-0.0078	-0.0068
Low income	-0.0106	-0.0134	-0.0087
AIC	27,338	28,660	21,396

* It is the best found solution.

4.6 Discussion

We used the Poisson regression to predict the number of STEMI patients in DAs, spatial units that were sufficiently small to provide useful travel time estimates. Our dependent variable was the number of STEMI patients, and our explanatory variables were the population in cohorts of age, sex, education, and income. We showed that the commonly used exponential link function had undesirable properties in our setting; we demonstrated that an identity link function of a linear, without intercept, combination of the explanatory variables was the only link function that was appropriate to predict the STEMI counts. We used the AIC statistic to compare

the predictive power of our Poisson regression model with that of standard multiple linear and negative binomial regressions against two sets of variables, namely reduced and full.

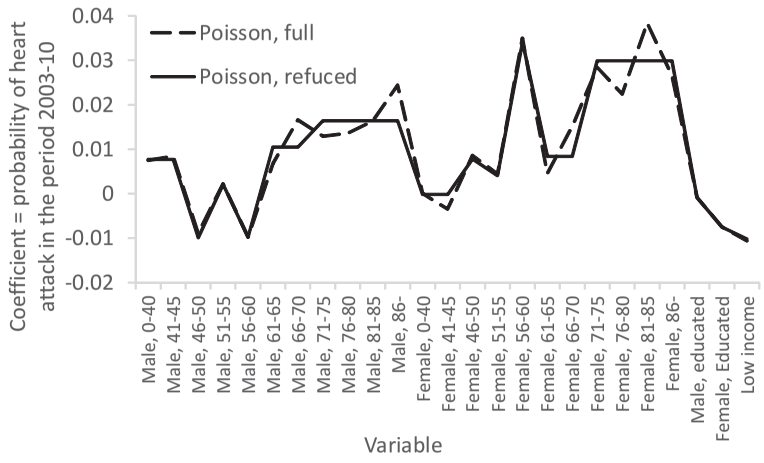
Interestingly, the AICs of all three regression approaches were similar across different data sets, but there was big differences among AICs across modeling approaches. Although negative binomial regression consistently had the minimum AIC in comparison with Poisson and linear regressions, Poisson regression was more favorable because calculations of the MLE estimators for Poisson regression were stable, but those calculations for negative binomial were unstable and did not return the optimal solutions. Two unexpected facts about results of our regression models required further research: (1) although each person has a non-zero probability of experiencing a heart attack, some of our coefficients were negative, implying that adding people in those cohorts would decrease the predicted number of heart attacks in a DA, and (2) the heart attack probability did not increase steadily by age.

Possible paths for investigating the non-intuitive patterns that we observed in our numerical results include: (1) Although Figure 4.5 showed that some of Poisson regression coefficients were consistently negative in 10 different sub-samples, one can use bootstrapping or other methods to compute confidence intervals for the coefficients to investigate whether the negative coefficients are significantly different from zero. One should consider that obtaining confidence intervals for coefficients of our count regression models is not an easy task, because of the constraints on the log-likelihood functions, (2) One might get insights into the negative coefficients by comparing DAs that are similar for all cohorts except one with a negative coefficient, and (3) One can investigate whether correlations between the health status of individuals, such as for married couples, could explain the non-intuitive patterns.

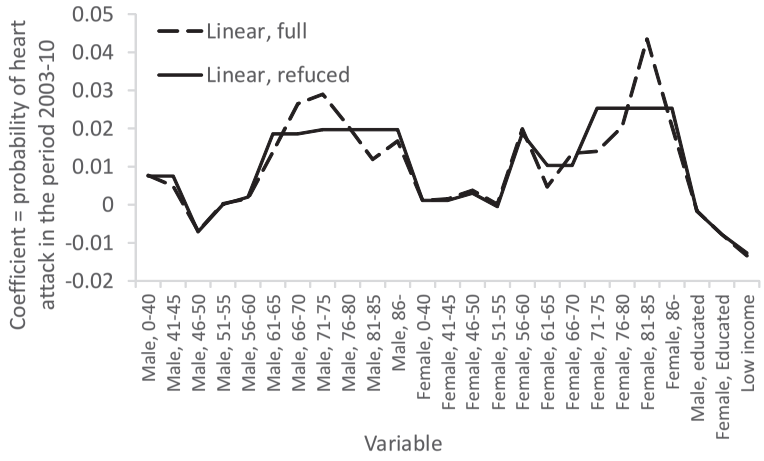
Possible directions for future research include: (1) developing a method, perhaps based on bootstrapping, to compute confidence intervals for the estimated coefficients, and (2) evaluating the impact of using each of the prediction methods discussed in this chapter on the PCI location problem and using the averages

of historical STEMI counts as a benchmark for prediction accuracy. To obtain a rough understanding of the performance of our Poisson model, we fitted a Poisson model to our 2003-2006 STEMI data and used that model to predict the number of STEMI patients for 2007-2010. As simpler benchmark methods, we also used the raw-average and last-point prediction methods to predict 2007-2010 STEMI incidence. For the former method, we divided the total number of 2003-2006 patients with the total population to obtain the heart attack chance of each person, and then multiplied that chance with each DA's population to obtain a prediction of the number of heart attack patients in that specific DA over 2007-2010, and for the latter method, we simply used the number of STEMI patients in each DA over 2003-2006 as a prediction for the number of heart attack patients in that DA over 2007-2010. We used the root-mean-square error (RMSE) to compare the performance of these three methods. The RMSE of the Poisson regression, raw-average, and last-point prediction methods were 2.17, 2.26, and 1.83, respectively.

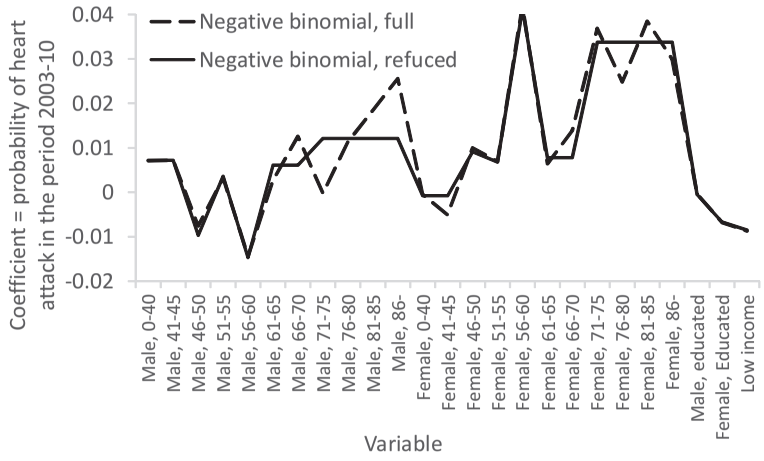
According to our data, the total number of STEMI patients during 2007-2010 were 4.7% more than that of 2003-2006. This increase could be because of an increase in the total population of Alberta or population aging, but we could not capture these changes with our method as we used the 2006 census data in both training and testing steps. That could be why our method did not perform as well as the last-point prediction method and its performance was just slightly better than that of the raw-average method.



(a) Poisson regression



(b) Multiple linear regression



(c) Negative binomial

Figure 4.6: Coefficients of reduced and full models are similar.

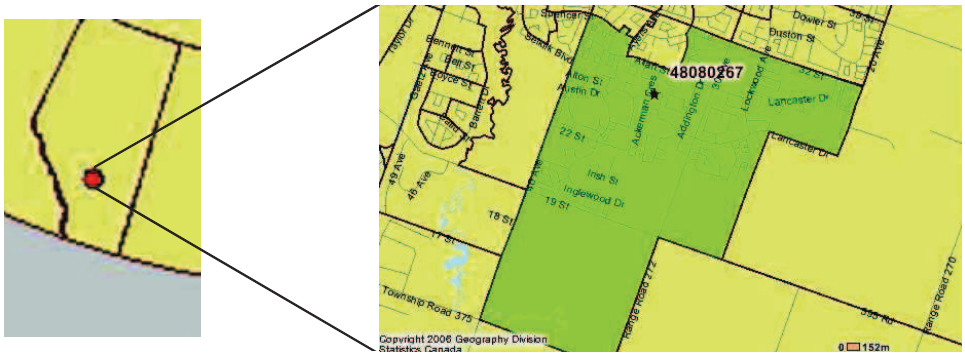


Figure 4.7: The province of Alberta and a close up of the most populous DA. *Note*. This figure is produced using the GeoSearch2006 product available at Statistics Canada (2009).

CHAPTER 5

Conclusion

In this dissertation we studied three separate problems in the heath care sector. In the first paper, presented in Chapter 2, we were interested in partial busy periods for loss systems because they corresponded to "Yellow Alert" and "Red Alert" periods for an ambulance system. During a Yellow Alert period, the number of available ambulances drops below some threshold, for example, 12 ambulances for the City of Calgary, while there is no ambulances available during a Red Alert. We introduced a recursive method to calculate the expected durations of partial busy periods for loss queueing systems, and showed that the expected durations of these periods were insensitive to the service time distribution beyond its mean. We used a big data set from Calgary EMS to validate our recursion. We also obtained the Laplace transform and moments of partial busy periods under different settings. Furthermore, we studied the impact of two actions, namely requesting new ambulances from neighboring cities and expediting the service of ambulances that are currently busy, on two performance measures, namely the duration of shortage periods and the number of lost calls. Possible future research includes model validation against the second moments, or standard deviations of empirical partial busy period durations, and investigating the impact of different service time distributions on the performance measures given a specific action.

In the second paper, presented in Chapter 3, we focused on the two main shortcomings of current computation approaches for stationary performance measures of

infinite-level LDQBDS. These shortcomings were: (1) the truncation level is chosen heuristically, and (2) it is not possible to calculate performance measures with a desired error tolerance. We addressed both of these issues by proposing two separate algorithms: In the first algorithm, we used Lyapunov analysis to find the truncation level for an infinite-level LDQBD such that the truncated upper tail is guaranteed to have less than a pre-specified proportion of the probability mass. In the second algorithm, we extended an algorithm from the literature such that the new updated algorithm truncates the state space automatically and calculates stationary performance measures with a desired accuracy. We use numerical examples to demonstrate both algorithms. Possible future research includes comparison of the two algorithms in terms of speed and accuracy, and extending the second algorithm to compute performance measures other than steady-state probabilities.

In the third paper, presented in Chapter 4, we proposed a tool to predict the number of heart attack patients in sufficiently small geographical units as a function of population size in cohorts of age, sex, education, and income. We used Poisson regression to build our model. We showed that the identity function of a linear combination of the explanatory variables, with no intercept, was the only functional form for the Poisson mean that satisfied properties of our heart attack counts and the commonly used exponential link function was not appropriate in this context. We built our models using an empirical data set of heart attack counts in Alberta postal codes from 2003 to 2010, and 2006 census data for dissemination areas. We used AIC to select our model variables and to compare the predictive power of our Poisson regression model with that of standard multiple linear and binomial regressions. Possible future research includes investigating the two unexpected observations that we made on our numerical results, namely the negative coefficients and the lack of monotonicity in coefficients with age, and applying our prediction models for finding good locations for heart attack treatment facilities.

BIBLIOGRAPHY

BIBLIOGRAPHY

Alanis, Ramon, Armann Ingolfsson, Bora Kolfal. 2013. A Markov chain model for an EMS system with repositioning. *Production and Operations Management* **22**(1) 216–231.

Albert, M. A., R. J. Glynn, J. Buring, P. M. Ridker. 2006. Impact of traditional and novel risk factors on the relationship between socioeconomic status and incident cardiovascular events. *Circulation* **114**(24) 2619–2626.

Almehdawe, E., B. Jewkes, Q. He. 2012. A Markovian queueing model for ambulance offload delays. *European Journal of Operational Research* .

American College of Emergency Physicians. 2013. Emergency Department wait times, crowding and access fact sheet URL `http://newsroom.acep.org/index.php?s=20301&item=29937`. Accessed June 6, 2015.

Armstrong, P. W., D. Collen, E. Antman. 2003. Fibrinolysis for acute myocardial infarction: The future is here and now. *Circulation* **107**(20) 2533–2537.

Artalejo, J. R., M. J. Lopez-Herrero. 2001. Analysis of the busy period for the $M/M/c$ queue: An algorithmic approach. *Journal of Applied Probability* **38**(1) 209–222.

Bakal, J. A., P. Kaul, R. C. Welsh, D. Johnstone, P. W. Armstrong. 2011. Determining the cost economic "tipping point" for the addition of a regional percutaneous coronary intervention facility. *Canadian Journal of Cardiology* **27**(5) 567–572.

Baker, D. W., C. D. Stevens, R. H. Brook. 1991. Patients who leave a public hospital emergency department without being seen by a physician: causes and consequences. *The Journal of the American Medical Association* **266**(8) 1085–1090.

Barrer, D. Y. 1957. Queuing with impatient customers and ordered service. *Operations Research* **5**(5) 650–656.

Baumann, H., W. Sandmann. 2012. Steady state analysis of level dependent quasi-birth-and-death processes with catastrophes. *Computers & Operations Research* **39**(2) 413–423.

Baumann, H., W. Sandmann. 2013. Computing stationary expectations in level-dependent QBD processes. *Journal of Applied Probability* **50**(1) 151–165.

Bazaraa, M. S., H. D. Sherali, C. M. Shetty. 2006. *Nonlinear Programming: Theory and Algorithms*. 3rd ed. Wiley-Interscience, Canada.

Berman, A., R. J. Plemmons. 1987. *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics.

Bose, S.K. 2013. *An Introduction to Queueing Systems*. Springer, U.S.

Botta, R. F., C. M. Harris. 1986. Approximation with generalized hyperexponential distributions: Weak convergence results. *Queueing Systems* **1**(2) 169–190.

Bountourelis, T., M. Y. Ulukus, J. P. Kharoufeh, S. G. Nabors. 2013. The Modeling, Analysis, and Management of Intensive Care Units. *Handbook of Healthcare Operations Management: Methods and Applications*. Springer, New York.

Brandt, A., M. Brandt. 2004. On the two-class $M/M/1$ system under preemptive resume and impatience of the prioritized customers. *Queueing Systems* **47**(1-2) 147–168.

Bright, L., P. G. Taylor. 1995. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Stochastic Models* **11**(3) 497–525.

Brumelle, S. L. 1978. A generalization of Erlang's loss system to state dependent arrival and service rates. *Mathematics of Operations Research* **3**(1) 10–16.

Campello, F., A. Ingolfsson, R. A. Shumsky. 2013. Queueing models of case managers. Working paper.

Carter, E. J., S. M. Pouch, E. L. Larson. 2014. The relationship between emergency department crowding and patient outcomes: a systematic review. *Journal of Nursing Scholarship* **46**(2) 106–115.

Cha, W. C., S. D. Shin, K. J. Song, S. K. Jung, G. J. Suh. 2009. Effect of an independent-capacity protocol on overcrowding in an urban emergency department. *Academic Emergency Medicine* **16**(12) 1277–1283.

Chan, C. W., V. F. Farias, G. Escobar. 2013. The impact of delays on service times in the intensive care unit. Working paper, Columbia Business School, Columbia University, New York, NY.

Chan, C. W., G. Yom-Tov, G. Escobar. 2011. When to use speedup: An examination of intensive care units with readmissions. *Operations Research* to appear.

Channouf, N., P. LEcuyer, A. Ingolfsson, A. N. Avramidis. 2007. The application of forecasting techniques to modeling emergency medical system calls in calgary, alberta. *Health Care Management Science* **10**(1) 25–45.

Choi, B. D., B. Kim, J. Chung. 2001. $M/M/1$ queue with impatient customers of higher priority. *Queueing Systems* **38**(1) 49–66.

Chow, V. S., M. L. Puterman, N. Salehirad, W. Huang, D. Atkins. 2011. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production and Operations Management* **20**(3) 418–430.

Cobham, A. 1954. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America* **2**(1) 70–76.

Cobham, A. 1955. Letter to the editor-Priority Assignment-a correction. *Journal of the Operations Research Society of America* **3**(4) 547–547.

Davis, R. H. 1966. Waiting-time distribution of a multi-server, priority queuing system. *Operations Research* **14**(1) 133–136.

Dayar, T., W. Sandmann, D. Spieler, V. Wolf. 2011. Infinite level-dependent QBD processes and matrix-analytic solutions for stochastic chemical kinetics. *Advances in Applied Probability* **43**(4) 1005–1026.

Delasay, M., A. Ingolfsson, B. Kolfal. 2013. Modeling load and overwork effects in queueing systems with adaptive service rates. Working paper.

Diez-Roux, A. V., B. G. Link, M. E. Northridge. 2000. A multilevel analysis of income inequality and cardiovascular disease risk factors. *Social science & medicine* **50**(5) 673–687.

Drekic, S., D. A. Stanford, D. G. Woolford, V. C. McAlister. 2015. A model for deceased-donor transplant queue waiting times. *Queueing Systems* **79**(1) 87–115.

Duran, A., G. Gutierrez, R. I. Zequeira. 2004. A continuous review inventory model with order expediting. *International Journal of Production Economics* **87**(2) 157–169.

Eitel, D. R., S. E. Rudkin, M. A. Malvehy, J. P. Killeen, J. M. Pines. 2010. Improving service quality by understanding emergency department flow: a white paper and position statement prepared for the American Academy of Emergency Medicine. *The Journal of Emergency Medicine* **38**(1) 70–79.

Erlander, S. 1967. A note on telephone traffic with losses. *Journal of Applied Probability* **4**(2) 406–408.

Fitch, J. J., R. A. Keller, D. Raynor, C. Zalar. 1993. *EMS Management*. JEMS, Carlsbad, CA.

Fogoros, R. N. 2008a. NSTEMI - Non ST Segment Myocardial Infarction. Accessed July 9, 2015, http://heartdisease.about.com/od/heartattack/g/NSTEMI.htm.

Fogoros, R. N. 2008b. STEMI - ST Segment Elevation Myocardial Infarction. Accessed July 9, 2015, http://heartdisease.about.com/od/heartattack/g/STEMI.htm.

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.

Gilboy, N., T. Tanabe, D. Travers, A. M. Rosenau. 2011. Emergency Severity Index (ESI): A triage tool for emergency department. *Rockville, MD: Agency for Healthcare Research and Quality* URL http://www.ahrq.gov/professionals/systems/hospital/esi/esi1.html. Accessed June 6, 2015.

Gordon, T., W. B. Kannel. 1982. Multiple risk functions for predicting coronary heart disease: the concept, accuracy, and application. *American heart journal* **103**(6) 1031–1039.

Goyal, A., D. L. Bhatt, P. G. Steg, B. J. Gersh, M. J. Alberts, E. M. Ohman, R. Corbalán, K. A. Eagle, E. Gaxiola, R. Gao, et al. 2010. Attained educational level and incident atherothrombotic events in low-and middle-income compared with high-income countriesclinical perspective. *Circulation* **122**(12) 1167–1175.

Green, L. V, P. J. Kolesar, J. Soares. 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research* **49**(4) 549–564.

Greene, W. H. 2012. *Econometric Analysis*. Prentice Hall.

Gross, D., C. M. Harris. 1998. *Fundamentals Of Queueing Theory (Wiley Series In Probability And Statistics)*. 3rd ed. Wiley-Interscience.

Grundy, S. M., D. Bilheimer, A. Chait, Luther T. C., M. Denke, R. J. Havel, W. R. Hazzard, S. B. Hulley, D. B. Hunninghake, R. A. Kreisberg, et al. 1993. Summary of the second report of the National Cholesterol Education Program (NCEP) Expert Panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel II). *Jama* **269**(23) 3015–3023.

Gu, W., X. Wang, S. E. McGregor. 2010. Optimization of preventive health care facility locations. *International Journal of Health Geographics* **9**(1) 17.

Henderson, S. G., A. J. Mason. 2004. Ambulance service planning: simulation and data visualisation. *Operations research and health care: a handbook of methods and applications* **70** 77–102.

Higginson, I. 2012. Emergency department crowding. *Emergency Medicine Journal* **29**(6) 437–443.

Hoot, N. R., D. Aronsky. 2008. Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of Emergency Medicine* **52**(2) 126–136.

Hulley, S., D. Grady, T. Bush, C. Furberg, D. Herrington, B. Riggs, E. Vittinghoff, et al. 1998. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Jama* **280**(7) 605–613.

Ignall, E., W. E. Walker. 1977. An analysis of the deployment of ambulances in Washington DC. *Journal of Urban Analysis* **4**(1) 59–92.

Ingolfsson, A., L. Tang. 2012. Efficient and reliable computation of birth-death process performance measures. *INFORMS Journal on Computing* **24**(1) 29–41.

Iravani, F., B. Balcıoğlu. 2008. On priority queues with impatient customers. *Queueing Systems* **58**(4) 239–260.

Izquierdo, L. R., S. S. Izquierdo, J. I. Santos, J. M. Galán, R. del Olmo. 2011. Teaching the mean-field approximation. *V international conference on industrial engineering and industrial management*. 502–506.

Jain, R., J. M. Smith. 1997. Modeling vehicular traffic flow using $M/G/C/C$ state dependent queueing models. *Transportation Science* **31**(4) 324–336.

Jouini, O., A. Roubos. 2014. Multiple priority multi-server queues. *Journal of the Operational Research Society* **65**(5) 616–632.

Jousilahti, P., E. Vartiainen, J. Tuomilehto, P. Puska. 1999. Sex, age, cardiovascular risk factors, and coronary heart disease a prospective follow-up study of 14 786 middle-aged men and women in finland. *Circulation* **99**(9) 1165–1172.

Kao, Edward P. C. 1996. *An Introduction to Stochastic Processes*. Cengage Learning, New York.

Kawanishi, K. 2008. Qbd approximations of a call center queueing model with general patience distribution. *Computers & Operations Research* **35**(8) 2463–2481.

KC, D. S., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50–65.

KC, Diwas. 2013. Does multitasking improve performance? Evidence from the emergency department. *Evidence from the Emergency Department (May 7, 2013)* .

Kella, O., U. Yechiali. 1985. Waiting times in the non-preemptive priority $M/M/c$ queue. *Stochastic Models* **1**(2) 257–262.

Kharoufeh, J. P. 2011. Level-dependent quasi-birth-and-death processes. *Wiley Encyclopedia of Operations Research and Management Science* .

Kim, S. H., M. A. Cohen, S. Netessine, S. Veeraraghavan. 2010. Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science* **56**(9) 1551–1567.

Kim, S.-H., W. Whitt. 2014. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing & Service Operations Management* forthcoming .

King, G. 1988. Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science* 838–863.

Kleinrock, L. 1964. A delay dependent queue discipline. *Naval Research Logistics Quarterly* **11**(3-4) 329–341.

Kleinrock, L., R. P. Finkelstein. 1967. Time dependent priority queues. *Operations Research* **15**(1) 104–116.

Kolesar, P. J., K. L. Rider, T. B. Crabill, W. E. Walker. 1975. A queuing-linear programming approach to scheduling police patrol cars. *Operations Research* **23**(6) 1045–1062.

Kutcher, M. A., L. W. Klein, F. S. Ou, T. P. Wharton Jr, G. J. Dehmer, M. Singh, H. V. Anderson, J. S. Rumsfeld, W. S. Weintraub, R. E. Shaw, et al. 2009. Percutaneous coronary interventions in facilities without cardiac surgery on site: A report from the National Cardiovascular Data Registry (NCDR). *Journal of the American College of Cardiology* **54**(1) 16.

Latouche, G., V. Ramaswami. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, Alexandria, Virginia.

Lawson, D. G., E. L. Porteus. 2000. Multistage inventory management with expediting. *Operations Research* **48**(6) 878–893.

Ledermann, W., G. E. H. Reuter. 1954. 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* 321–369.

Li, A. A., W. Whitt. 2013. Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. *Performance Evaluation* forthcoming .

Lin, D., J. Patrick, F. Labeau. 2014. Estimating the waiting time of multi-priority emergency patients with downstream blocking. *Health care management science* **17**(1) 88–99.

Luenberger, D. G. 1979. *Introduction to Dynamic Systems: Theory, Models, and Applications*. Wiley, New York.

Marschner, I.C. 2010. Stable computation of maximum likelihood estimates in identity link Poisson regression. *Journal of Computational and Graphical Statistics* **19**(3) 666–683.

Martins, D., K. Nelson, D. Pan, N. Tareen, K. Norris. 2001. The effect of gender on age-related blood pressure changes and the prevalence of isolated systolic hypertension among older adults: Data from NHANES III. *The Journal of Gender-specific Medicine* **4**(3) 10–13.

Mason, A. J. 2013. Simulation And Real-Time Optimised Relocation For Improving Ambulance Operations. *Handbook of Healthcare Operations Management*. Springer, 289–317.

Maxwell, M. S., M. Restrepo, S. G. Henderson, H. Topaloglu. 2010. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing* **22**(2) 266–281.

Alberta Health Services. 2010. Emergency department surge capacity protocols. Accessed November 17, 2012, `http://www.albertahealthservices.ca/3167.asp`.

Canada Post. 2012. Canada Post Glossary. Accessed July 9, 2015, `http://www.canadapost.ca/tools/pg/glossary-e.asp#1`.

Heart & Stroke Foundation. 2013. The Heart and Stroke Foundation Statistics. Accessed July 9, 2015, `http://www.heartandstroke.ab.ca/site/c.lqIRL1PJJtH/b.3650897/k.35F8/Statistics.htm#heartattack`.

Knitro. 2015. Knitro-Matlab. Accessed July 9, 2015, `http://www.ziena.com/knitromatlab.htm`.

Statistics Canada. 2009. GeoSearch 2006. Accessed July 9, 2015, `http://geodepot.statcan.ca/GeoSearch2006/GeoSearch2006.jsp?resolution=H&lang=E&otherLang=F`.

Statistics Canada. 2010. 2006 Census Dictionary. Accessed July 9, 2015, `http://www12.statcan.gc.ca/census-recensement/2006/ref/dict/index-eng.cfm`.

Stony Brook University Medical Center. 2012. Transferring admitted emergency department patients to hallway beds leads to lower length of stay and higher patient satisfaction. Accessed November 7, 2013, `http://www.innovations.ahrq.gov/content.aspx?id=2840`.

The College of Emergency Medicine. 2012. Crowding in emergency departments. Accessed November 7, 2013, `http://www.collemergencymed.ac.uk/Shop-Floor/Clinical%20Guidelines/College%20Guidelines/default.asp`.

McHugh, M., P. Tanabe, M. McClelland, R. K. Khare. 2012. More patients are triaged using the Emergency Severity Index than any other triage acuity system in the United States. *Academic Emergency Medicine* **19**(1) 106–109.

Meyn, S. P., R. L. Tweedie. 1993. Stability of markovian processes iii: Foster-Lyapunov criteria for continuous-time processes. *Advances in Applied Probability* 518–548.

Miller, D. R. 1981a. Computation of steady-state probabilities for $M/M/1$ priority queues. *Operations Research* **29**(5) 945–958.

Miller, K. S. 1981b. On the inverse of the sum of matrices. *Mathematics Magazine* 67–72.

myFOXdetroit. 2011. No ambulance available for shooting victims. *myFOXdetroit* 10 April. Accessed November 17, 2012, `http://www.myfoxdetroit.com/story/18473311/no-ambulance-available-for-shooting-victims`.

Nelson, R. 1995. *Probability, Stochastic Processes, And Queueing Theory: The Mathematics Of Computer Performance Modelling*. Springer, New York.

Neuts, M. F. 1981. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Corporation.

Omahen, K., V. Marathe. 1978. Analysis and applications of the delay cycle for the $M/M/c$ queueing system. *Journal of the ACM* **25**(2) 283–303.

Oredsson, S., H. Jonsson, J. Rognes, L. Lind, K. E. Goransson, A. Ehrenberg, K. Asplund, M. Castrén, N. Farrohknia. 2011. A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* **19**(1) 43.

Palm, C. 1957. *Research on telephone traffic carried by full availability groups*. Tele.

Patel, A. B., N. M. Waters, W. A. Ghali. 2007. Determining geographic areas and populations with timely access to cardiac catheterization facilities for acute myocardial infarction care in Alberta, Canada. *International Journal of Health Geographics* **6** 47.

Patrick, J., K. Nelson, D. Lane. 2015. A simulation model for capacity planning in community care. *Journal of Simulation* **9**(2) 111–120.

Pedersen, L. 2012. Too few paramedics to answer call: Union official. *Toronto Sun* 13 May. Accessed November 17, 2012, `http://www.torontosun.com/2012/05/13/too-few-paramedics-to-answer-call-union-of\unhbox\voidb@x\hbox{}ficial`.

Phung-Duc, Tuan, Hiroyuki Masuyama, Shoji Kasahara, Yutaka Takahashi. 2010. A simple algorithm for the rate matrices of level-dependent qbd processes. *Proceedings of the 5th International Conference on Queueing Theory and Network Applications*. ACM, 46–52.

Pines, J. M., R. J. Batt, J. A. Hilton, C. Terwiesch. 2011. The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Annals of Emergency Medicine* **58**(4) 331–340.

Restrepo, M., S. G. Henderson, H. Topaloglu. 2009. Erlang loss models for the static deployment of ambulances. *Health care management science* **12**(1) 67–79.

Rosengren, A., S. V. Subramanian, S. Islam, C. K. Chow, A. Avezum, K. Kazmi, K. Sliwa, M. Zubaid, S. Rangarajan, S. Yusuf. 2009. Education and risk for acute myocardial infarction in 52 high, middle and low-income countries: INTERHEART case-control study. *Heart* **95**(24) 2014–2022.

Rowe, B. H., P. Channan, M. Bullard, S. Blitz, L. D. Saunders, R. J. Rosychuk, H. Lari, W. R. Craig, B. R. Holroyd. 2006. Characteristics of patients who leave emergency departments without being seen. *Academic Emergency Medicine* **13**(8) 848–852.

Rozenshmidt, Lubov. 2008. On priority queues with impatient customers: Stationary and time-varying analysis. Ph.D. thesis, Master's thesis, Technion, Israel Institute of Technology.

Sanchez, M., A. J. Smally, R. J. Grant, L. M. Jacobs. 2006. Effects of a fast-track area on emergency department performance. *The Journal of Emergency Medicine* **31**(1) 117–120.

Sarhangian, V., B. Balcıoğlu. 2013. Waiting time analysis of multi-class queues with impatient customers. *Probability in the Engineering and Informational Sciences* **27**(03) 333–352.

Schiff, J. 1999. *The Laplace Transform: Theory And Applications*. Springer, New York.

Schmid, V. 2012. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research* **219**(3) 611–621.

Schneider, K. 2012. Opposition demands EMS wait time review. *Calgary Sun* 24 February. Accessed November 17, 2012, `http://www.calgarysun.com/2012/02/24/opposition-demands-ems-wait-time-review`.

Setzler, H., C. Saydam, S. Park. 2009. EMS call volume predictions: A comparative study. *Computers & Operations Research* **36**(6) 1843–1851.

Sharma, O. 1990. *Markovian Queues*. Ellis Horwood, New York.

Shmueli, G. 2010. To explain or to predict? *Statistical Science* **25**(3) 289–310.

Sinnema, J. 2010. ER waits keep paramedics away from calls: Average 12 minutes longer in hallways waiting for patients to be seen. *Edmonton Journal* 28 October. Accessed November 17, 2012, `http://www2.canada.com/edmontonjournal/news/cityplus/story.html?id=c265934a-267e-40a4-a5e3-2b4872f9fd65`.

Sinnema, J. 2012. Busy ambulance system causes concern for paramedics. *Edmonton Journal* 20 January. Accessed November 17, 2012, `http://www2.canada.com/edmontonjournal/news/archives/story.html?id=0a795570-199f-4ed0-91a5-c1c562743faa`.

Stanford, D. A., P. Taylor, I. Ziedins. 2014. Waiting time distributions in the accumulating priority queue. *Queueing Systems* **77**(3) 297–330.

Steutel, F. W. 1970. Preservation of infinite divisibility under mixing and related topics. *MC Tracts* **33** 1–99.

Sun, B. C., R. Y. Hsia, R. E. Weiss, D. Zingmond, L.-J. Liang, W. Han, H. McCreath, S. M. Asch. 2013. Effect of emergency department crowding on outcomes of admitted patients. *Annals of Emergency Medicine* **61**(6) 605–611.

Sun, X. 2008. Modeling the emergency care delivery system using a queueing approach. Ph.D. thesis, Master's thesis, University of Waterloo.

Takács, L. 1969. On Erlang's formula. *The annals of mathematical statistics* **40**(1) 71–78.

Taylor, P.G. 2013. Insensitivity In Stochastic Models. *Queueing Networks: A Fundamental Approach (International Series in Operations Research & Management Science)*. Springer, New York, 121–140.

Veeraraghavan, S., A. Scheller-Wolf. 2008. Now or later: A simple policy for effective dual sourcing in capacitated systems. *Operations Research* **56**(4) 850–864.

Wang, J., O. Baron, A. Scheller-Wolf. 2015. $M/M/c$ queue with two priority classes. *Operations Research* .

Watase, T., R. Fu, D. Foster, D. Langley, D. A. Handel. 2012. The impact of an ED-only full-capacity protocol. *The American journal of emergency medicine* **30**(8) 1329–1335.

Weiss, A., L. Williams, J. M. Smith. 2012. Performance & optimization of $M/G/c/c$ building evacuation networks. *Journal of Mathematical Modelling and Algorithms* **11**(4) 361–386.

Whitt, W. 2012. Fitting birth-and-death queueing models to data. *Statistics & Probability Letters* **82**(5) 998–1004.

Wielgosz, A., M. Arango, C. Bancej, A. Bienek, H. Johansen, P. Lindsay, W. Luo, A. Luteyn, Cyril N., P. Quan, et al. 2009. Tracking heart disease and stroke in Canada 2009. Tech. Rep. HP32-3/2009E, Public Health Agency of Canada.

Wilson, P. W. F., R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, W. B. Kannel. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation* **97**(18) 1837–1847.

Winkleby, M. A., D. E. Jatulis, E. Frank, S. P. Fortmann. 1992. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *American journal of public health* **82**(6) 816–820.

Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice Hall.

Yildiz, Ozlem, Michael F Kamali, Tolga Tezcan. 2015. Analysis of triage systems in emergency departments. *Working Paper. Available at SSRN 2617687* .

Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue. *Queueing Systems* **51**(3-4) 361–402.

Zhang, Z. G., H. P. Luh, C.-H. Wang. 2011. Modeling security-check queues. *Management Science* **57**(11) 1979–1995.

APPENDICES

APPENDIX A

Proofs and Notations for Modeling Yellow and Red Alert Durations

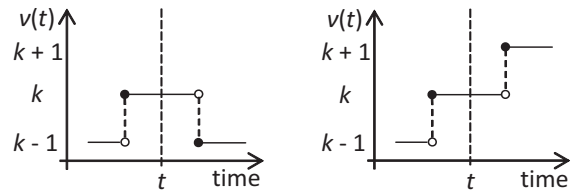
A.1 Section 2.3 Proofs

We first prove Theorem 2.3, then Theorem 2.2, and then Theorem 2.1 because we refer to Theorem 2.3 and its proof when we prove Theorems 2.1 and 2.2. We use $f_X(x)$, $F_X(x)$, and $\overline{F}_X(x)$ to denote the PDF, CDF, and complementary CDF of a positive and continuous random variable X, respectively, $\mathrm{E}(X)$ for the expected value, $\tilde{X} = (X - t|X > t)$ for the residual lifetime, $\mathscr{L}_X(s)$ for the LT of the PDF of X, and $\mathscr{L}_{f(y)}(s)$ to denote the LT of a function $f(y)$: $\mathscr{L}_{f(y)}(s) = \int_0^\infty e^{-sy} f(y)\mathrm{d}y$.

A.1.1 Theorem 2.3 Proof

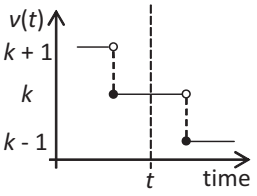
Proof. Recursion (2.2) holds for the $M/G/c/c$ system:

We begin by proving $\mathrm{E}(B_c) = 1/(c\mu)$. Let $\{E_1, E_2, \dots\}$ and $\{\phi_1, \phi_2, \dots\}$ be the sequence of events (arrivals or departures) and the sequence of corresponding epochs, respectively. If the current time is t, we define the events $N_{\nu(t)}$ ("next event is an arrival") and $L_{\nu(t)}$ ("last event was an arrival") and their complements as

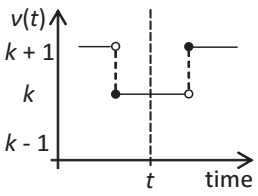


(a) L_k and N_k^c

(b) L_k and N_k



(c) L_k^c and N_k^c



(d) L_k^c and N_k

Figure A.1: Possible combinations of the last and next events when $\nu(t) = k$.

follows:

$$N_{\nu(t)} = \{E_n \text{ is an arrival} \mid n = \min\{i : \phi_i > t\}\},$$
$$N^c_{\nu(t)} = \{E_n \text{ is a departure} \mid n = \min\{i : \phi_i > t\}\},$$
$$L_{\nu(t)} = \{E_n \text{ is an arrival} \mid n = \max\{i : \phi_i \leq t\}\},$$
$$L^c_{\nu(t)} = \{E_n \text{ is a departure} \mid n = \max\{i : \phi_i \leq t\}\}.$$

Figure A.1 shows all possible combinations of these events. We use $R_{\nu(t)} = \phi_{i+1} - \phi_i$, where $t \in [\phi_i, \phi_{i+1})$ to denote the time between two consecutive events.

Suppose that a c-partial busy period begins at t_0, which means that an arrival occurs at t_0 and $\nu(t_0) = c$. Since all c servers are busy at t_0, the next event after t_0 must be a departure, which occurs at $t_0 + R_{\nu(t_0)}$ leaving $c - 1$ busy servers behind, which means that the c-partial busy period ends at $t_0 + R_{\nu(t_0)}$. Therefore,

$$\mathrm{E}(B_c) = \mathrm{E}(B_{\nu(t_0)}|L_{\nu(t_0)}) = \mathrm{E}(R_{\nu(t_0)}|L_{\nu(t_0)}) = \mathrm{E}(R_c|L_c). \tag{A.1}$$

The first equality in (A.1) is a reminder that a c-partial busy period always begins with a customer arrival. The third equality follows because there are c busy servers

at t_0.

Given L_c, it follows from (2.1) that R_c is the minimum of c independent random variables, specifically $c-1$ residual service times and one service time, that is,

$$B_c = R_c|L_c = \min\left\{T, \tilde{T}_1, \ldots, \tilde{T}_{c-1}\right\}. \tag{A.2}$$

Therefore,

$$\overline{F}_{B_c}(t) = \overline{F}_{R_c|L_c}(t) = \Pr\{T > t, \tilde{T}_1 > t, \ldots, \tilde{T}_{c-1} > t\} = \overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{c-1}, \tag{A.3}$$

$$\mathrm{E}(B_c) = \mathrm{E}(R_c|L_c) = \int_0^\infty \overline{F}_{R_c|L_c}(t)\,\mathrm{d}t = \int_0^\infty \overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{c-1}\,\mathrm{d}t. \tag{A.4}$$

In order to simplify (A.4), we define $Y_k(t) = \overline{F}_{\tilde{T}}(t)^k$ and use $F_{\tilde{T}}(t) = \mu\int_0^t \overline{F}_T(s)\,\mathrm{d}s$. Then

$$\frac{\mathrm{d}\,Y_k(t)}{\mathrm{d}\,t} = \frac{\mathrm{d}\left(1-\mu\int_0^t \overline{F}_T(s)\,\mathrm{d}s\right)^k}{\mathrm{d}\,t} = -k\mu\overline{F}_T(t)Y_{k-1}(t) \Rightarrow \tag{A.5}$$

$$\frac{\mathrm{d}\,Y_c(t)}{\mathrm{d}\,t} = -c\mu\overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{c-1}. \tag{A.6}$$

By comparing the integrand in (A.4) with (A.6), we see that (A.4) simplifies as follows:

$$\mathrm{E}(B_c) = -\frac{1}{c\mu}\int_0^\infty \mathrm{d}\,\overline{F}_{\tilde{T}}(t)^c = -\frac{1}{c\mu}\left(\overline{F}_{\tilde{T}}(\infty)^c - \overline{F}_{\tilde{T}}(0)^c\right) = \frac{1}{c\mu}, \tag{A.7}$$

which completes the proof of $\mathrm{E}(B_c) = 1/(c\mu)$.

To prove the recursion $\mathrm{E}(B_k) = (\lambda\mathrm{E}(B_{k+1}))/(k\mu) + 1/(k\mu)$, we assume that a k-partial busy period begins at t_0, where $k \in \{1, \ldots, c-1\}$, which means that an arrival occurred at t_0 and $\nu(t_0) = k$. We decompose the duration B_k into (1) the time from t_0 until the next event epoch t_1 and (2) the time from t_1 until the epoch

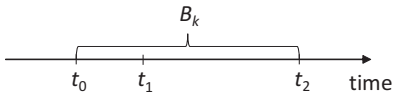


Figure A.2: A schematic view of B_k and its components when $t_2 - t_1 > 0$.

t_2 when the k-partial busy period ends, as illustrated in Figure A.2, so that

$$\mathrm{E}(B_k) = \mathrm{E}(t_1 - t_0) + \mathrm{E}(t_2 - t_1). \tag{A.8}$$

If the event at t_1 is a departure, with probability $\Pr(N^c_{\nu(t_0)}|L_{\nu(t_0)})$, then the number of busy ambulances drops to $k-1$ at t_1 and the k-partial busy period ends; that is, $\mathrm{E}(B_k) = \mathrm{E}(t_1 - t_0)$. Otherwise, if the event at t_1 is an arrival, with probability $\Pr(N_{\nu(t_0)}|L_{\nu(t_0)})$, then the number of busy ambulances increases to $k+1$ at t_1 and a $(k+1)$-partial busy period begins; that is, $\mathrm{E}(t_2 - t_1) > 0$ and $\mathrm{E}(B_k) = \mathrm{E}(t_1 - t_0) + \mathrm{E}(t_2 - t_1)$. Therefore,

$$\begin{aligned}\mathrm{E}(B_k) = \mathrm{E}(B_{\nu(t_0)}|L_{\nu(t_0)}) &= \Pr(N^c_{\nu(t_0)}|L_{\nu(t_0)})\mathrm{E}(t_1 - t_0|L_{\nu(t_0)}) \\ &+ \Pr(N_{\nu(t_0)}|L_{\nu(t_0)})\left(\mathrm{E}(t_1 - t_0|L_{\nu(t_0)}) + \mathrm{E}(t_2 - t_1|N_{\nu(t_0)}, L_{\nu(t_0)})\right)\end{aligned} \tag{A.9}$$

Hereafter, for convenience, we replace $\nu(t_0)$ with its value k.

The time interval $t_1 - t_0$ is equal to $R_k|L_k$. Given L_k, R_k is the minimum of $k+1$ independent random variables, specifically $k-1$ residual service times, one service time, and one inter-arrival time, that is,

$$t_1 - t_0 = R_k|L_k = \min\left\{Q, T, \tilde{T}_1, \ldots, \tilde{T}_{k-1}\right\}. \tag{A.10}$$

Therefore,

$$\overline{F}_{R_k|L_k}(t) = \Pr\{Q > t, T > t, \tilde{T}_1 > t, \ldots, \tilde{T}_{k-1} > t\} = e^{-\lambda t}\overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{k-1}, \tag{A.11}$$

$$\mathrm{E}(t_1 - t_0|L_k) = \mathrm{E}(R_k|L_k) = \int_0^\infty \overline{F}_{R_k|L_k}(t)\,\mathrm{d}t = \int_0^\infty e^{-\lambda t}\overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{k-1}\,\mathrm{d}t. \tag{A.12}$$

By using (A.5), (A.12) can be represented as

$$\begin{aligned}\mathrm{E}(R_k|L_k) &= \int_0^\infty e^{-\lambda t}\overline{F}_T(t)Y_{k-1}(t)\,\mathrm{d}t = -\frac{1}{k\mu}\int_0^\infty\left(e^{-\lambda t}\frac{\mathrm{d}\,Y_k(t)}{\mathrm{d}\,t}\right)\mathrm{d}t \\ &= -\frac{1}{k\mu}(\lambda\mathscr{L}_{Y_k(t)}(\lambda)-1). \end{aligned} \tag{A.13}$$

Equation (A.13) results from the property $\mathscr{L}_{\frac{\mathrm{d}\,f(t)}{\mathrm{d}\,t}}(s) = s\mathscr{L}_{f(t)}(s) - f^+(0)$ (Schiff 1999, p. 209).

Next, we calculate $\mathrm{E}(t_2 - t_1|N_k, L_k)$. If L_k and N_k, then there is at least one $(k+1)$-partial busy period nested inside the current k-partial busy period. When a nested $(k+1)$-partial busy period ends, the next change in $\nu(t)$ could be either an increase or a decrease. If $\nu(t)$ increases to $k+1$, with probability $\Pr(N_k|L_k^c)$, then another $(k+1)$-partial busy period begins. Otherwise, if $\nu(t)$ decreases to $k-1$, with probability $\Pr(N_k^c|L_k^c)$, then the k-partial busy period ends. If we consider the completion of the k-partial busy period as a "success," then the number of times that the system enters a $(k+1)$-partial busy period before the k-partial busy period ends follows a geometric distribution with parameter $\Pr(N_k^c|L_k^c)$. Therefore, $\mathrm{E}(t_2 - t_1|N_k, L_k)$ consists of a random number $N \in \{1, 2, \dots\}$ of cycles, each cycle with duration $R_k|L_k^c + B_{k+1}$. As $\mathrm{E}(N) = 1/\Pr(N_k^c|L_k^c)$, we have

$$\mathrm{E}(t_2 - t_1|N_k \cap L_k) = \frac{1}{\Pr(N_k^c|L_k^c)}\left(\mathrm{E}(R_k|L_k^c) + \mathrm{E}(B_{k+1})\right). \tag{A.14}$$

Substituting (A.12) and (A.14) in (A.9) results in:

$$\mathrm{E}(B_k) = \mathrm{E}(R_k|L_k) + \frac{\Pr(N_k|L_k)}{\Pr(N_k^c|L_k^c)}(\mathrm{E}(R_k|L_k^c) + \mathrm{E}(B_{k+1})) \tag{A.15}$$

Given L_k^c, the random variable R_k is the minimum of $k+1$ independent random variables, specifically, k residual service times and one inter-arrival time, that is,

$$R_k|L_k^c = \min\left\{Q, \tilde{T}_1, \dots, \tilde{T}_k\right\}. \tag{A.16}$$

Therefore,

$$\overline{F}_{R_k|L_k^c}(t) = \Pr\{Q > t, \tilde{T}_1 > t, \ldots, \tilde{T}_k > t\} = e^{-\lambda t}\overline{F}_{\tilde{T}}(t)^k, \tag{A.17}$$

$$\mathrm{E}(R_k|L_k^c) = \int_0^\infty \overline{F}_{R_k|L_k^c}(t)\,\mathrm{d}t = \int_0^\infty e^{-\lambda t}\overline{F}_{\tilde{T}}(t)^k\,\mathrm{d}t = \mathscr{L}_{Y_k(t)}(\lambda). \tag{A.18}$$

We obtain the last equality in (A.18) using the definition $Y_k(t) = \overline{F}_{\tilde{T}}(t)^k$.

The only remaining unknowns in (A.15) are $\Pr(N_k|L_k)$ and $\Pr\left(N_k^c|L_k^c\right)$. These two probabilities can be calculated by conditioning on the inter-arrival time as follows:

$$\begin{aligned}
\Pr(N_k|L_k) &= \int_0^\infty f_Q(t)\Pr(N_k|Q=t, L_k)\,\mathrm{d}t \\
&= \int_0^\infty \lambda e^{-\lambda t}\Pr\{T > t, \tilde{T}_1 > t, \ldots, \tilde{T}_{k-1} > t\}\,\mathrm{d}t \\
&= \int_0^\infty \lambda e^{-\lambda t}\overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{k-1}\,\mathrm{d}t, \qquad \text{(A.19)}
\end{aligned}$$

$$\begin{aligned}
1 - \Pr(N_k^c|L_k^c) = \Pr(N_k|L_k^c) &= \int_0^\infty f_Q(t)\Pr(N_k|Q=t, L_k^c)\,\mathrm{d}t \\
&= \int_0^\infty \lambda e^{-\lambda t}\Pr\{\tilde{T}_1 > t, \ldots, \tilde{T}_k > t\}\,\mathrm{d}t = \int_0^\infty \lambda e^{-\lambda t}\overline{F}_{\tilde{T}}(t)^k\,\mathrm{d}t. \qquad \text{(A.20)}
\end{aligned}$$

Comparing (A.19) with (A.12) and (A.20) with (A.18), we obtain

$$\Pr(N_k|L_k) = \lambda\mathrm{E}(R_k|L_k), \tag{A.21}$$

$$\Pr(N_k|L_k^c) = \lambda\mathrm{E}(R_k|L_k^c). \tag{A.22}$$

Using (A.21)-(A.22) in (A.15), we obtain:

$$\mathrm{E}(B_k) = \frac{\mathrm{E}(R_k|L_k)\left(1 + \lambda\mathrm{E}(B_{k+1})\right)}{1 - \lambda\mathrm{E}(R_k|L_k^c)}. \tag{A.23}$$

By using (A.13) and the last equality of (A.18) in (A.23), we obtain (2.2).

□

Proof. Recursion (2.2) holds for the $M/D/c/c$ system:

With deterministic service times, we have

$$F_T(t) = \begin{cases} 1 & \text{if } t \geq 1/\mu \\ 0 & \text{otherwise} \end{cases} \tag{A.24}$$

$$F_{\tilde{T}}(t) = \mu \int_0^t \overline{F}_T(s)\,\mathrm{d}s = \begin{cases} 1 & \text{if } t \geq 1/\mu \\ \mu t & \text{if } 0 \leq t < 1/\mu \end{cases} \tag{A.25}$$

Therefore,

$$\mathrm{E}(B_c) = \int_0^\infty \overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{c-1}\,\mathrm{d}t = \int_0^{1/\mu} (1-\mu t)^{c-1}\,\mathrm{d}t = -\frac{1}{\mu}\int_1^0 x^{c-1}dx = \frac{1}{c\mu}. \tag{A.26}$$

When the service time is deterministic, (A.26) shows that (A.4) reduces to the first equation of (2.2). Next, we use (A.15), (A.27), and (A.32) to show that the second equation of (2.2) also holds when the service time is deterministic. Note that (A.15) holds even if the service time is not continuous.

We use induction to show that

$$k\mu\mathrm{E}(R_k|L_k) + \lambda\mathrm{E}(R_k|L_k^c) = 1, \quad k = 1, \ldots, c-1, \tag{A.27}$$

where, following the same logic as (A.12) and (A.18), we have

$$\mathrm{E}(R_k|L_k) = \int_0^\infty e^{-\lambda t}\overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{k-1}\,\mathrm{d}t = \int_0^{1/\mu} e^{-\lambda t}(1-\mu t)^{k-1}\,\mathrm{d}t, \tag{A.28}$$

$$\mathrm{E}(R_k|L_k^c) = \int_0^\infty e^{-\lambda t}\overline{F}_{\tilde{T}}(t)^{k}\,\mathrm{d}t = \int_0^{1/\mu} e^{-\lambda t}(1-\mu t)^{k}\,\mathrm{d}t. \tag{A.29}$$

For $k = 1$, (A.28)-(A.29) reduce to (A.27). As the induction hypothesis, we assume that (A.27) holds for $k \in \{1, ..., c-2\}$, and finish the proof by showing that (A.27)

then holds for $k+1$. We use integration by parts ($\int u\mathrm{d}v = uv - \int v\mathrm{d}u$), and represent (A.28)-(A.29) as:

$$\mathrm{E}(R_k|L_k) = \frac{1}{\lambda} - \frac{(k-1)\mu}{\lambda}\,\mathrm{E}(R_{k-1}|L_{k-1}), \tag{A.30}$$

$$\mathrm{E}(R_k|L_k^c) = \frac{1}{\lambda} - \frac{k\mu}{\lambda}\,\mathrm{E}(R_{k-1}|L_{k-1}^c). \tag{A.31}$$

If we multiply both sides of equations (A.30)-(A.31) by $k\mu$ and λ, respectively, and add them up, then we observe that (A.27) also holds for $k+1$ if it holds for $k \in \{1, ..., c-2\}$.

As (A.21)-(A.22) hold even if the service time distribution is not continuous, we obtain (A.32) from (A.27).

$$\frac{k\mu}{\lambda}\Pr(N_k|L_k) + \Pr(N_k|L_k^c) = 1, \quad k = 1, \ldots, c-1. \tag{A.32}$$

Using (A.27) and (A.32) in (A.15), we complete the proof. □

Proof. The higher moments of B_c are sensitive to the distribution of T: We know that the following equation holds if X is a non-negative continuous random variable and n is a positive integer (Wolff 1989, p. 37).

$$\mathrm{E}(X^n) = \int_0^\infty \overline{F}_X(u^{1/n})\,\mathrm{d}u. \tag{A.33}$$

When $n = 1$, we have the well-known $\mathrm{E}(X) = \int_0^\infty \overline{F}_X(u)\,\mathrm{d}u$. Combining (A.3) and (A.33), we obtain:

$$\mathrm{E}(B_c^n) = \int_0^\infty \overline{F}_{B_c}(u^{1/n})\,\mathrm{d}u = \int_0^\infty \overline{F}_T(u^{1/n})\overline{F}_{\tilde{T}}(u^{1/n})^{c-1}\,\mathrm{d}u. \tag{A.34}$$

We use a counter example to complete the proof: If $c = 1$, then

$$\mathrm{E}(B_c^n) = \begin{cases} n!/0.5^n & \text{if } T \text{ has an exponential distribution with mean 2} \\ 4^n/(1+n) & \text{if } T \text{ has a uniform distribution over [0,4].} \end{cases} \tag{A.35}$$

From this example, we see that $\mathrm{E}(B_c^n)$, for $c = 1$ and $n \geq 2$, depends on the distribution of T. □

Proof. The LTs $\mathscr{L}_{B_k}(s)$ for $M/G/c/c$ systems satisfy (2.8)-(2.9): Following the same line of reasoning we used to obtain (A.1), we get $B_c = R_c|L_c$. Therefore, $f_{B_c}(t) = f_{R_c|L_c}(t)$, and $\mathscr{L}_{B_c}(s) = \mathscr{L}_{R_c|L_c}(s)$ which proves (2.8). To prove (2.9), we follow the line of reasoning we used to obtain (A.9). As $B_k = R_k|L_k$ with probability $\Pr(N_k^c|L_k)$, and $B_k = R_k|L_k + \sum_{i=1}^{N}(B_{k+1}^i + R_k^i|L_k^c)$ with probability $\Pr(N_k|L_k)$, where $B_{k+1}^i + R_k^i|L_k^c$ are i.i.d. random variables for $i = 1, \ldots, N$. That is,

$$\begin{aligned} f_{B_k}(t) &= \Pr(N_k^c|L_k) f_{R_k|L_k}(t) \\ &\quad + \Pr(N_k|L_k) f_{R_k|L_k + \sum_{i=1}^{N}(B_{k+1}^i + R_k^i|L_k^c)}(t), \quad k = c-1, \ldots, 1. \end{aligned} \tag{A.36}$$

If we take the LT of both sides of (A.36) for $k = c-1, \ldots, 1$, we obtain

$$\mathscr{L}_{B_k}(s) = \Pr(N_k^c|L_k)\mathscr{L}_{R_k|L_k}(s) + \Pr(N_k|L_k)\mathscr{L}_{R_k|L_k}(s)\mathscr{L}_{\sum_{i=1}^{N}(B_{k+1}^i + R_k^i|L_k^c)}(s). \tag{A.37}$$

From (Nelson 1995, p. 200), if $X_1 + \cdots + X_N$ is a random sum of i.i.d. random variables with the common LT $\mathscr{L}_X(s)$, then,

$$\mathscr{L}_{\sum_{i=1}^{N} X_i}(s) = \mathscr{L}_X(s)^N = \mathrm{E}_N\left(\mathscr{L}_X(s)^N\right) = \mathscr{G}_N\left(\mathscr{L}_X(s)\right), \tag{A.38}$$

where $\mathscr{G}_N(s)$ is the probability-generating function (PGF) of N. As discussed in Section A.1.1, the random variable N in (A.37) is geometrically distributed with parameter $\Pr(N_k^c|L_k^c)$ and support $\{1, 2, \ldots\}$, which implies that the PGF of N is

$\mathscr{G}_N(s) = Pr(N_k^c|L_k^c)s/(1 - \Pr(N_k|L_k^c)s)$. The random variables $B_{k+1}^i + R_k^i|L_k^c$ are i.i.d., and therefore,

$$\mathscr{L}_{\sum_{i=1}^N (B_{k+1}^i + R_k^i|L_k^c)}(s) = \mathscr{G}_N\left(\mathscr{L}_{B_{k+1}}(s)\mathscr{L}_{R_k|L_k^c}(s)\right) \tag{A.39}$$

$$= \frac{\Pr(N_k^c|L_k^c)\mathscr{L}_{B_{k+1}}(s)\mathscr{L}_{R_k|L_k^c}(s)}{1 - \Pr(N_k|L_k^c)\mathscr{L}_{B_{k+1}}(s)\mathscr{L}_{R_k|L_k^c}(s)}. \tag{A.40}$$

Substituting (A.40) in (A.37) we obtain (2.9). □

The unknown components of (2.9) are $\mathscr{L}_{R_k|L_k}(s)$ and $\mathscr{L}_{R_k|L_k^c}(s)$. Applying the property $\mathscr{L}_{\frac{\mathrm{d}\, f(t)}{\mathrm{d}\, t}}(s) = s\mathscr{L}_{f(t)}(s) - \lim_{t\to 0^+} f(t)$, we have $\mathscr{L}_{R_k|L_k}(s) = s\mathscr{L}_{F_{R_k|L_k}(t)}(s)$. Therefore,

$$\begin{aligned}\mathscr{L}_{R_k|L_k}(s) &= s\mathscr{L}_{F_{R_k|L_k}(t)}(s) = s\mathscr{L}_{1-\overline{F}_{R_k|L_k}(t)}(s)\\ &= s\int_0^\infty e^{-st}\left(1 - e^{-\lambda t}\overline{F}_T(t)\overline{F}_{\tilde{T}}(t)^{k-1}\right)\,\mathrm{d}t\\ &= 1 - \frac{s}{k\mu} + \frac{s(\lambda+s)}{k\mu}\mathscr{L}_{Y_k(t)}(\lambda+s).\end{aligned} \tag{A.41}$$

The third equality in (A.41) follows from (A.11) and the last one follows from (A.13). Similar to $\mathscr{L}_{R_k|L_k}(s)$, we calculate $\mathscr{L}_{R_k|L_k^c}(s)$ using (A.17) and (A.18) :

$$\mathscr{L}_{R_k|L_k^c}(s) = s\int_0^\infty e^{-st}\left(1 - e^{-\lambda t}\overline{F}_{\tilde{T}}(t)^k\right)\,\mathrm{d}t = 1 - s\mathscr{L}_{Y_k(t)}(\lambda+s). \tag{A.42}$$

A.1.2 Theorem 2.2 Proof

Random variable X has generalized hyperexponential (GH) distribution with m components if the distribution of X is a mixture of m exponential distributions, that is,

$$F_X(x) = 1 - \sum_{i=1}^m a_i e^{-\lambda_i x}, \text{ where } \sum_{i=1}^m a_i = 1. \tag{A.43}$$

The expected value and the LT of X are:

$$\mathrm{E}(X) = \sum_{i=1}^{m} \frac{a_i}{\lambda_i}, \ \mathscr{L}_X(s) = \sum_{i=1}^{m} a_i \frac{\lambda_i}{s+\lambda_i}. \tag{A.44}$$

If all of the weights on the exponential components of X are positive, then X has a hyperexponential (H) distribution.

Random variable X has a generalized hyper-Erlang (GHE) distribution with m components if the distribution of X is a mixture of m Erlang distributions, that is,

$$\mathscr{L}_X(s) = \sum_{i=1}^{m} a_i \left(\frac{\lambda_i}{s+\lambda_i}\right)^{m_i}, \ \sum_{i=1}^{m} a_i = 1. \tag{A.45}$$

If all of the shape parameters m_i are equal to 1, then X has a GH distribution.

We stress that if $i \neq j$ in (A.43), then $\lambda_i \neq \lambda_j$, but we can have $\lambda_i = \lambda_j$ for distinct i and j in (A.45), if $m_i \neq m_j$. The class of H distributions are a proper subset of GH distributions, and GH distributions are a proper subset of GHE distributions.

If X and Y are two independent GH random variables with weights a_i and b_j, and rates λ_i and μ_j for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, respectively, then the following results hold:

Lemma A.1. *The residual of X, $\tilde{X}$, has a GH distribution with the same rates as X, but different weights.*

Proof.

$$\begin{aligned} F_{\tilde{X}}(x) &= \frac{1}{\mathrm{E}(X)} \int_0^x \left(\sum_{i=1}^{m} a_i e^{-\lambda_i t} \right) \mathrm{d}t = \frac{1}{\mathrm{E}(X)} \sum_{i=1}^{m} \left(\frac{a_i}{\lambda_i} - \frac{a_i}{\lambda_i} e^{-\lambda_i x} \right) \\ &= 1 - \frac{1}{\mathrm{E}(X)} \sum_{i=1}^{m} \frac{a_i}{\lambda_i} e^{-\lambda_i x}. \end{aligned} \tag{A.46}$$

The last equation holds because

$$\frac{1}{\mathrm{E}(X)} \sum_{i=1}^{m} \frac{a_i}{\lambda_i} = \frac{\mathrm{E}(X)}{\mathrm{E}(X)} = 1. \tag{A.47}$$

Comparing (A.46) with (A.43), $F_{\tilde{X}}(x)$ is the CDF of a *GH* distribution if the coefficients of (A.46) add up to unity, which is true as follows from (A.47). □

Lemma A.2. *The minimum of X and Y has a GH distribution with at most mn components.*

Proof. Let $Z = \min\{X, Y\}$, then,

$$\begin{aligned} F_Z(z) &= 1 - \Pr(Z > z) = 1 - \left(\sum_{i=1}^{m} a_i e^{-\lambda_i z}\right)\left(\sum_{j=1}^{n} b_j e^{-\mu_j z}\right) \\ &= 1 - \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j e^{-(\lambda_i+\mu_j)z}\right) \end{aligned} \tag{A.48}$$

Comparing (A.48) with (A.43), $F_Z(z)$ is the CDF of a *GH* distribution if the coefficients of (A.48) add up to unity, which is true:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j = \sum_{i=1}^{m} a_i \sum_{j=1}^{n} b_j = 1. \tag{A.49}$$

The number of components for Z equals the number of distinct values for $\lambda_i + \mu_j$, which is mn or less.

□

Lemma A.3. *The weighted sum $\alpha_1 \mathscr{L}_{X_1}(s) + \cdots + \alpha_n \mathscr{L}_{X_n}(s)$ is the LT of a GHE random variable when the weights, α_j, add up to unity, and X_j are independent (but not necessarily identical) GHE random variables with $\mathscr{L}_{X_j}(s) = \sum_{i=1}^{m_j} a_{ij} \left(\frac{\lambda_{ij}}{s+\lambda_{ij}}\right)^{m_{ij}}$, where $\sum_{i=1}^{m_j} a_{ij} = 1$, for $j = 1, ..., n$.*

Proof. Let $Z(s) = \alpha_1 \mathscr{L}_{X_1}(s) + \cdots + \alpha_n \mathscr{L}_{X_n}(s)$, where $\sum_{j=1}^{n} \alpha_j = 1$, then,

$$Z(s) = \sum_{j=1}^{n} \alpha_j \mathscr{L}_{X_j}(s) = \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{m_j} a_{ij} \left(\frac{\lambda_{ij}}{s+\lambda_{ij}}\right)^{m_{ij}}. \tag{A.50}$$

Comparing (A.50) with (A.45), $Z(s)$ is the LT of a *GHE* distribution if the coefficients of (A.50) add up to unity, which is true: $\sum_{j=1}^{n} \alpha_j \sum_{i=1}^{m_j} a_{ij} = 1$.

□

Lemma A.4. *The sum of two independent GHE random variables has a GHE distribution.*

Proof. Before we prove the lemma, we obtain two preliminary results by using induction. First, we use induction to show that if $\mathscr{L}_X(s) = \lambda/(s+\lambda)$ and $\mathscr{L}_Y(s) = (\mu/(s+\mu))^n$, then, by Lemma A.3, $X+Y$ is GHE with rates λ and μ, and the shape parameter associated with λ is 1. If $n = 1$, then

$$\mathscr{L}_{X+Y}(s) = \mathscr{L}_X(s)\mathscr{L}_Y(s) = \left(\frac{\lambda}{s+\lambda}\right)\left(\frac{\mu}{s+\mu}\right) = a\left(\frac{\lambda}{s+\lambda}\right) + b\left(\frac{\mu}{s+\mu}\right), \tag{A.51}$$

where $a = \mu/(\mu-\lambda)$ and $b = -\lambda/(\mu-\lambda)$. As $a+b=1$, Lemma A.3 is satisfied. Assume as an induction hypothesis that the statement holds when n is an arbitrary positive integer. Given this assumption, we prove that the statement holds when the shape parameter of Y is $n+1$. We assume that

$$\left(\frac{\lambda}{s+\lambda}\right)\left(\frac{\mu}{s+\mu}\right)^n = \alpha\frac{\lambda}{s+\lambda} + \sum_{j=1}^{l} b_j \left(\frac{\mu}{s+\mu}\right)^{l_j}, \tag{A.52}$$

where l and l_j are positive integers, and $\alpha + \sum_{j=1}^{l} b_j = 1$. Multiplying both sides of (A.52) with $\mu/(s+\mu)$ and using (A.51), we obtain

$$\left(\frac{\lambda}{s+\lambda}\right)\left(\frac{\mu}{s+\mu}\right)^{n+1} = \alpha a\left(\frac{\lambda}{s+\lambda}\right) + \alpha b\left(\frac{\mu}{s+\mu}\right) + \sum_{j=1}^{l} b_j \left(\frac{\mu}{s+\mu}\right)^{l_j+1}. \tag{A.53}$$

The weights on the components of the right hand side of (A.53) add up to unity. Therefore, by Lemma A.3, $X+Y$ is GHE with LT:

$$\mathscr{L}_{X+Y}(s) = \mathscr{L}_X(s)\mathscr{L}_Y(s) = \alpha a\left(\frac{\lambda}{s+\lambda}\right) + \alpha b\left(\frac{\mu}{s+\mu}\right) + \sum_{j=1}^{l} b_j \left(\frac{\mu}{s+\mu}\right)^{l_j+1}. \tag{A.54}$$

Second, we use induction to show that if $\mathscr{L}_X(s) = (\lambda/(s+\lambda))^m$ and $\mathscr{L}_Y(s) = (\mu/(s+\mu))^n$, then $X+Y$ is GHE with rates λ and μ. If at least one of m or n equals 1, then the statement holds as shown in (A.51) and (A.54). Assume as an induction hypothesis that the statement holds when m and n are arbitrary positive integers. Given this assumption, we prove that the statement holds when the shape parameter of X is $m+1$ and that of Y is n. (The case in which the shape parameter of X is m and that of Y is $n+1$ can be proven similarly.) We assume that

$$\left(\frac{\lambda}{s+\lambda}\right)^m \left(\frac{\mu}{s+\mu}\right)^n = \sum_{i=1}^{k} a_i \left(\frac{\lambda}{s+\lambda}\right)^{k_i} + \sum_{j=1}^{l} b_j \left(\frac{\mu}{s+\mu}\right)^{l_j}. \tag{A.55}$$

where k, l, k_i, and l_j are positive integers, and $\sum_{i=1}^{k} a_i + \sum_{j=1}^{l} b_j = 1$. Multiplying both sides of (A.55) with $\lambda/(s+\lambda)$, we obtain

$$\left(\frac{\lambda}{s+\lambda}\right)^{m+1} \left(\frac{\mu}{s+\mu}\right)^n = \sum_{i=1}^{k} a_i \left(\frac{\lambda}{s+\lambda}\right)^{k_i+1} + \sum_{j=1}^{l} b_j \left(\frac{\lambda}{s+\lambda}\right) \left(\frac{\mu}{s+\mu}\right)^{l_j}. \tag{A.56}$$

Following (A.54), (A.56) is a weighted sum of the LTs of GHE random variables with weights that add up to unity. Therefore, (A.56) is the LT of a GHE random variable (Lemma A.3) and the statement is proven.

Now, we prove that the sum of two independent GHE random variables is GHE. If

$$\mathscr{L}_X(s) = \sum_{i=1}^{m} a_i \left(\frac{\lambda_i}{s+\lambda_i}\right)^{m_i} \text{ and } \mathscr{L}_Y(s) = \sum_{j=1}^{n} b_j \left(\frac{\mu_j}{s+\mu_j}\right)^{n_j},$$

then

$$\mathscr{L}_{X+Y}(s) = \mathscr{L}_X(s)\mathscr{L}_Y(s) = \left(\sum_{i=1}^{m} a_i \left(\frac{\lambda_i}{s+\lambda_i}\right)^{m_i}\right)\left(\sum_{j=1}^{n} b_j \left(\frac{\mu_j}{s+\mu_j}\right)^{n_j}\right)$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \left(\frac{\lambda_i}{s+\lambda_i}\right)^{m_i}\left(\frac{\mu_j}{s+\mu_j}\right)^{n_j} = \sum_{i=1}^{m}\sum_{j|\mu_j=\lambda_i} a_i b_j \left(\frac{\lambda_i}{s+\lambda_i}\right)^{m_i+n_j}$$
$$+ \sum_{i=1}^{m}\sum_{j|\mu_j\neq\lambda_i} a_i b_j \left(\frac{\lambda_i}{s+\lambda_i}\right)^{m_i}\left(\frac{\mu_j}{s+\mu_j}\right)^{n_j}. \tag{A.57}$$

In the last term of (A.57), for each i and j, $((\lambda_i)/(s+\lambda_i))^{m_i}((\mu_j)/(s+\mu_j))^{n_j}$ is the LT of a GHE random variable (it follows from (A.56)). Therefore, (A.57) is a weighted sum of the LTs of GHE random variables with weights that add up to unity. That is, (A.57) is the LT of a GHE random variable (Lemma A.3) and the lemma is proven. □

Lemma A.5. *If N is a positive-integer-valued random variable with PGF $\mathscr{G}_N(s)$, then $\mathscr{G}_N(\mathscr{L}_X(s))$ is the LT of a GHE distribution.*

Proof. By the definition of a PGF,

$$\mathscr{G}_N(\mathscr{L}_X(s)) = \mathrm{E}_N\left(\mathscr{L}_X(s)^N\right) = \sum_{n=1}^{\infty} \Pr(N=n)\mathscr{L}_X(s)^n. \tag{A.58}$$

In (A.58), $\mathscr{L}_X(s)^n$ is the LT of the sum of n i.i.d. random variables $X_1, \ldots, X_n$, which has a GHE distribution (Lemma A.4). Therefore, $\mathscr{G}_N(\mathscr{L}_X(s))$ is a weighted sum of the LTs of GHE random variables with weights that add up to unity, which is the LT of a GHE random variable according to Lemma A.3.

□

Now, we have all material we need to prove Theorem 2.2.

Proof. Proof of Theorem 2.2.

We use induction, starting with the index c as the base case (instead of 1) and work backwards in the index set $\{c-1, ..., 1\}$. We first show that B_c is GHE. The service time T is GH, which implies that the residual service time $\tilde{T}$ is also GH

(Lemma A.1). It means that, according to (A.2), B_c is the minimum of a set of GH random variables, which, following Lemma A.2, also has a GH distribution. Since GH is a proper subset of GHE, B_c is GHE.

Assume as an induction hypothesis that B_{k+1} is GHE. Given this assumption, we prove that B_k is GHE:

Substituting (A.39) in (A.37), we obtain the LT of B_k as:

$$\begin{aligned}\mathscr{L}_{B_k}(s) =& \Pr(N_k^c|L_k)\mathscr{L}_{R_k|L_k}(s) \\ &+ \Pr(N_k|L_k)\mathscr{L}_{R_k|L_k}(s)\mathscr{G}_N\left(\mathscr{L}_{B_{k+1}}(s)\mathscr{L}_{R_k|L_k^c}(s)\right),\ k=c-1,\ldots,1.\end{aligned} \tag{A.59}$$

According to (A.10), (A.16), and Lemma A.1, $R_k|L_k$ and $R_k|L_k^c$ are the minimums of GH random variables. Therefore, following Lemma A.2, $R_k|L_k$ and $R_k|L_k^c$ have GH distributions. Since both B_{k+1} and $R_k|L_k^c$ are GHE, $\mathscr{L}_{B_{k+1}}(s)\mathscr{L}_{R_k|L_k^c}(s)$ is the LT of a GHE random variable (Lemma A.4). Therefore, following Lemma A.5, $\mathscr{G}_N\left(\mathscr{L}_{B_{k+1}}(s)\mathscr{L}_{R_k|L_k^c}(s)\right)$ is the LT of a GHE. Applying Lemma A.4 again shows that $\mathscr{L}_{R_k|L_k}(s)\mathscr{G}_N\left(\mathscr{L}_{B_{k+1}}(s)\mathscr{L}_{R_k|L_k^c}(s)\right)$ is the LT of a GHE. Therefore, the LT of B_k in (A.59) is a weighted sum of the LTs of two GHE random variables with weights that add up to unity and B_k is GHE (Lemma A.3).

□

A.1.3 Theorem 2.1 Proof

Proof. The recursion, closed form, and the monotonicity of $\mathrm{E}(B_k)$: We proved the recursion (2.2) for a more general system in Appendix A.1.1. The closed form solution (2.6) for $\mathrm{E}(B_k)$ is obtained by successively replacing $\mathrm{E}(B_{k+1})$ with its equation in (2.2) starting with $\mathrm{E}(B_c)$. The recursion (2.2) implies that $\mathrm{E}(B_{c-1}) > \mathrm{E}(B_c)$. In general, (2.6) implies that $\mathrm{E}(B_k) > \mathrm{E}(B_{k+1})$—one can confirm the inequality as

follows:

$$\mathrm{E}(B_k) - \mathrm{E}(B_{k+1}) = \frac{1}{\mu} \sum_{i=0}^{c-k-1} \frac{(k-1)!}{(k+i)!} \left(1 - \frac{k}{(k+i+1)}\right) \left(\frac{\lambda}{\mu}\right)^i + \frac{(k-1)!}{\mu\, c!} \left(\frac{\lambda}{\mu}\right)^{c-k} > 0, \quad k \in \{1, \ldots, c-2\}. \tag{A.60}$$

□

The remainder of the proof proceeds as follows. We prove (2.5) next and use it to prove (2.3) and (2.7). The recursion (2.4) follows directly from (2.2) and (2.3).

Proof. In $M/M/c/c$ systems, B_k satisfy (2.5) and follow an H distribution with $c-k+1$ components: Due to the memoryless property of the exponential distribution, the components of (2.8)-(2.9) from Theorem 3 are independent of the last event, so we drop conditioning on L and L^c in this section.

The properties of the exponential distribution, combined with (2.9), imply that:

$$\mathscr{L}_{B_c}(s) = \frac{c\mu}{s + c\mu}, \tag{A.61}$$

$$\Pr(N_k) = \frac{\lambda}{\lambda + k\mu}, \ \mathscr{L}_{R_k}(s) = \frac{\lambda + k\mu}{s + \lambda + k\mu}, \ k = 1, ..., c-1,$$

$$\mathscr{L}_{B_k}(s) = \frac{k\mu}{\lambda + k\mu + s - \lambda \mathscr{L}_{B_{k+1}}(s)}, \ k = c-1, \ldots, 1. \tag{A.62}$$

To show that B_k follows an H distribution, we define the function $P_m(s)$ as:

$$P_m(s) = \begin{cases} 1 & \text{if } m = 0 \\ c\mu + s & \text{if } m = 1 \\ (\lambda + (c-m+1)\mu + s)P_{m-1}(s) - \lambda(c-m+2)\mu P_{m-2}(s) & \text{if } m \geq 2 \end{cases} \tag{A.63}$$

We can represent $\mathscr{L}_{B_{c-1}}(s)$ in (A.62) using $P_1(s)$ and $P_2(s)$:

$$\mathscr{L}_{B_{c-1}}(s) = \frac{(c-1)\mu P_1(s)}{P_2(s)}. \tag{A.64}$$

We obtain the following expression for $\mathscr{L}_{B_{c-m}}(s)$ by recursively using (A.62):

$$\mathscr{L}_{B_{c-m}}(s) = \frac{(c-m)\mu P_m(s)}{P_{m+1}(s)}, \quad m = 1, \ldots, c-1. \tag{A.65}$$

Lemma 1 in Ledermann and Reuter (1954) implies that the roots of $P_m(s) = 0$, which we will denote as $-r_1^{(m)}, ..., -r_m^{(m)}$ are distinct and negative, for $m = 1, ..., c$. The same lemma implies that the roots of $P_m(s) = 0$ are separated by the roots of $P_{m-1}(s) = 0$. Therefore, we can represent $P_m(s)$ as:

$$P_m(s) = \prod_{i=1}^{m} \left(s + r_i^{(m)}\right), \tag{A.66}$$

such that

$$0 < r_1^{(m)} < r_1^{(m-1)} < r_2^{(m)} < r_2^{(m-1)} < \cdots < r_{m-1}^{(m-1)} < r_m^{(m)}, \ m = 2, ..., c. \tag{A.67}$$

From (A.65) and (A.66), we obtain:

$$\mathscr{L}_{B_{c-m}}(s) = (c-m)\mu \frac{\prod_{i=1}^{m}\left(s + r_i^{(m)}\right)}{\prod_{i=1}^{m+1}\left(s + r_i^{(m+1)}\right)}, \quad m = 1, \ldots, c-1. \tag{A.68}$$

It follows from (A.67), (A.68), and Lemma 2.12.1 in Steutel (1970) that there are unique $p_1 > 0, ..., p_{m+1} > 0$ such that

$$\mathscr{L}_{B_{c-m}}(s) = (c-m)\mu \frac{\prod_{i=1}^{m}\left(s + r_i^{(m)}\right)}{\prod_{i=1}^{m+1}\left(s + r_i^{(m+1)}\right)} = \sum_{i=1}^{m+1} p_i \frac{r_i^{(m+1)}}{s + r_i^{(m+1)}}, \quad m = 1, \ldots, c-1, \tag{A.69}$$

where $\sum_{i=1}^{m+1} p_i = 1$. The right hand side of (A.69) is the LT of an H random variable with $m+1$ components. If $k = c - m$, then $\mathscr{L}_{B_k}(s)$ is the LT of an H random variable with $c - k + 1$ components. □

Proof. The recursion, closed form, and monotonicity of $\mathrm{Var}(B_k)$: The nth moment,

$\mathrm{E}(X^n)$, of a continuous random variable X can be obtained from its LT as follows:

$$\mathrm{E}(X^n) = (-1)^n \mathscr{L}_X^{(n)}(0), \tag{A.70}$$

where $\mathscr{L}_X^{(n)}(s) = (d^n/ds^n)\mathscr{L}_X(s)$. We use (A.70) to obtain the second moment of B_k:

$$\mathrm{E}(B_c^2) = \mathscr{L}_{B_c}^{(2)}(0) = \left.\frac{2c\mu}{(s+c\mu)^3}\right|_{s=0} = 2\mathrm{E}(B_c)^2, \tag{A.71}$$

$$\begin{aligned}\mathrm{E}(B_k^2) = \mathscr{L}_{B_k}^{(2)}(0) &= \left.\frac{\lambda}{k\mu}\mathscr{L}_{B_{k+1}}^{(2)}(s)\right|_{s=0} + 2\left(\frac{1}{k\mu} - \frac{\lambda}{k\mu}\mathscr{L}_{B_{k+1}}^{(1)}(s)\right)^2\Bigg|_{s=0} \\ &= \frac{\lambda}{k\mu}\mathrm{E}(B_{k+1}^2) + 2\mathrm{E}(B_k)^2,\ k = c-1,\ldots,1,\end{aligned} \tag{A.72}$$

where $\mathrm{E}(B_c)$ and $\mathrm{E}(B_k)$ can be obtained from (2.2) or (2.6).

Following (A.71)-(A.72), the variance of B_k, $\mathrm{Var}(B_k) = \mathrm{E}(B_k^2) - \mathrm{E}(B_k)^2$, is:

$$\mathrm{Var}(B_c) = \mathrm{E}(B_c)^2, \tag{A.73}$$

$$\mathrm{Var}(B_k) = \frac{\lambda}{k\mu}\mathrm{E}(B_{k+1}^2) + \mathrm{E}(B_k)^2. \tag{A.74}$$

We add and subtract $(\lambda/(k\mu))\mathrm{E}(B_{k+1})^2$ to (A.74), and obtain the recursion (2.3). If we successively replace $\mathrm{Var}(B_{k+1})$ with its equation in (2.3), we obtain the closed form solution for $\mathrm{Var}(B_k)$ in (2.7) .

If we successively replace $\mathrm{E}(B_{k+1}^2)$ in (A.72) starting from $k = c$ and working backwards in $\{c-1, ..., 1\}$ we obtain:

$$\mathrm{E}(B_k^2) = 2\sum_{i=0}^{c-k}\left(\frac{\lambda}{\mu}\right)^i \frac{(k-1)!}{(k-1+i)!}\mathrm{E}(B_{k+i})^2,\ k = c-1,\ldots,1, \tag{A.75}$$

which implies that

$$\begin{aligned}\mathrm{E}(B_k^2) - \mathrm{E}(B_{k+1}^2) = 2\sum_{i=0}^{c-k-1}\frac{(k-1)!}{(k+i)!}\left((k+i)\mathrm{E}(B_{k+i})^2 - k\mathrm{E}(B_{k+i+1})^2\right)\left(\frac{\lambda}{\mu}\right)^i \\ + 2\frac{(k-1)!}{(c-1)!}\left(\frac{\lambda}{\mu}\right)^{c-k}\mathrm{E}(B_c)^2, \quad k = 1,\ldots,c-1.\end{aligned} \tag{A.76}$$

As shown earlier in this section, $\mathrm{E}(B_k)$ decreases as k increases. Therefore, (A.76) implies that $\mathrm{E}(B_k^2)$ is also decreasing with k. As a result,

$$\mathrm{Var}(B_k) - \mathrm{Var}(B_{k+1}) = \mathrm{E}(B_k^2) - \mathrm{E}(B_{k+1}^2) + \mathrm{E}(B_{k+1})^2 - \mathrm{E}(B_k)^2 > 0,$$
$$k = 1, \ldots, c-1, \tag{A.77}$$

so the variance of B_k decreases as k increases. Note that the inside-summation expression in (A.76) reduces to $\mathrm{E}(B_k)^2 - \mathrm{E}(B_{k+1})^2$ for $i = 0$, which cancels out the only negative part of (A.77), $\mathrm{E}(B_{k+1})^2 - \mathrm{E}(B_k)^2$.

□

A.2 Heuristic for Section 2.4

We explain how we segmented weekdays. We used a similar heuristic for weekends. We pool all weekdays and estimated the arrival rate $\hat{\lambda}_i$ and scheduled number of ambulances $\hat{c}_i$ for each hour $i = 0, 1, ..., 23$. First, we create Segment 1, consisting of hours $0, ..., u_0$, where u_0 is the largest integer that satisfies $\max_{0\le i\le u_0} \hat{\lambda}_i - \min_{0\le i\le u_0} \hat{\lambda}_i \le \epsilon_\lambda$ and $\max_{0\le i\le u_0} \hat{c}_i - \min_{0\le i\le u_0} \hat{c}_i \le \epsilon_c$, where ϵ_λ and ϵ_c are our tolerance levels for variation in the arrival rate and scheduled number of ambulances, respectively. Second, we create Segment 2, consisting of hours $u_0 + 1, ..., u_1$, where u_1 is the largest integer that satisfies $\max_{u_0+1\le i\le u_1} \hat{\lambda}_i - \min_{u_0+1\le i\le u_1} \hat{\lambda}_i \le \epsilon_\lambda$ and $\max_{u_0+1\le i\le u_1} \hat{c}_i - \min_{u_0+1\le i\le u_1} \hat{c}_i \le \epsilon_c$. We continue creating segments in this way until all of the 24 hours are assigned to a segment.

A.3 Figures for Section 2.4

A.4 Section 2.5 Proof

A.4.1 Theorem 2.4 Proof

Proof. The random variable $\tilde{B}_{ii}|L_i^c$ is the time from the moment that one of nested $(i+1)$-partial busy periods within the i-partial busy period ends to the end of the i-partial busy period. Therefore, $\tilde{B}_{ii}|L_i^c$ equals the time to finish a service after a

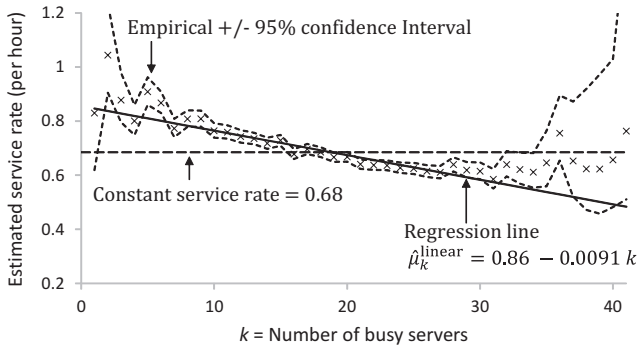


Figure A.3: Average service rate as a function of the number of busy ambulances.

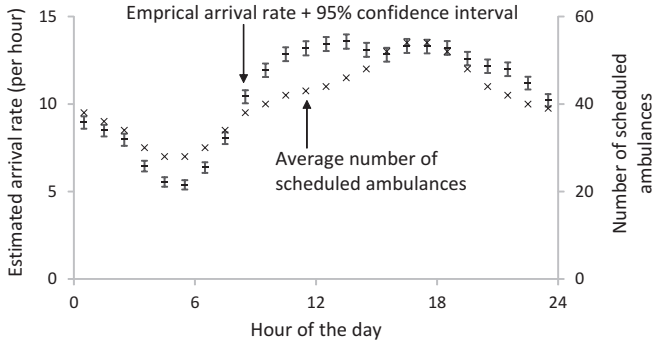


Figure A.4: The arrival rate and number of scheduled ambulances by time.

nested $(i+1)$-partial busy period ends, $R_i|L_i^c$, plus a geometrically distributed random number $N' \in \{0, 1, \dots\}$ of cycles, each with duration $B_{i+1} + R_i|L_i^c$. Therefore,

$$\mathrm{E}\left(\tilde{B}_{ii}|L_i^c\right) = \mathrm{E}(R_i|L_i^c) + \left(\frac{1}{\Pr(N_i^c|L_i^c)} - 1\right)(\mathrm{E}(R_i|L_i^c) + \mathrm{E}(B_{k+1})). \tag{A.78}$$

We use $\Pr(N_i|L_i^c) = 1 - \Pr(N_i^c|L_i^c)$ to represent (A.78) as a function of $\Pr(N_i|L_i^c)$, and then apply (A.22) to obtain (2.19).

We proved the expected sojourn time formula, (A.18), as a part of the proof of Theorem 2.3 in Appendix A.1.1. □

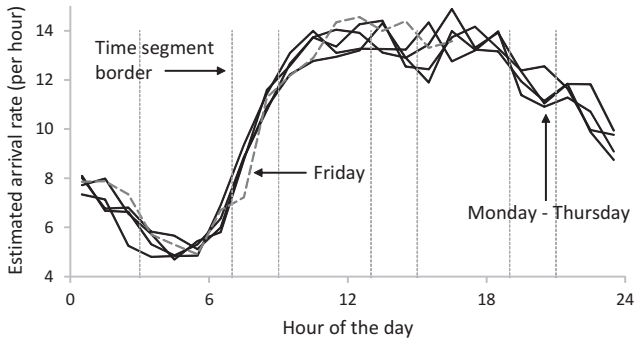


Figure A.5: Weekdays: 00:00 on Monday to 19:00 on Friday.

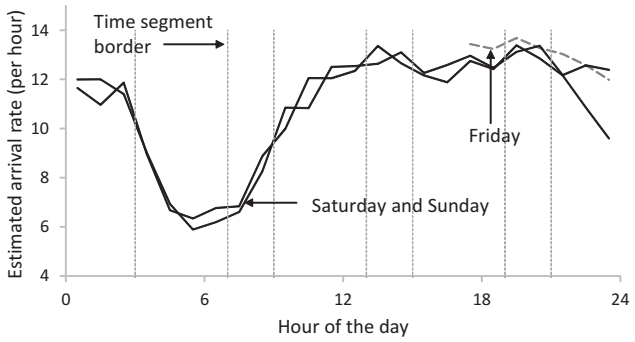


Figure A.6: Weekend: 19:00 on Friday to 24:00 on Sunday.

A.5 Expected Sojourn Times for Section 2.5.2.2

For transient state s, we obtain $\mathrm{E}(R_s)$ as:

$$\mathrm{E}\left(R_{(c,1,0)}\right) = \mathrm{E}\left(\min\left\{V, T, \tilde{T}_1, \ldots, \tilde{T}_{c-1}\right\}\right) = \int_0^\infty e^{-\delta r}\overline{F}_T(r)\overline{F}_{\tilde{T}}(r)^{c-1}\,\mathrm{d}r, \tag{A.79}$$

$$\mathrm{E}\left(R_{(m,1,0)}\right) = \mathrm{E}\left(\min\left\{Q, V, T, \tilde{T}_1, \ldots, \tilde{T}_{m-1}\right\}\right) = \int_0^\infty e^{-(\lambda+\delta)r}\overline{F}_T(r)\overline{F}_{\tilde{T}}(r)^{m-1}\,\mathrm{d}r,$$

$$m = k+1, \ldots, c-1, \tag{A.80}$$

$$\mathrm{E}\left(R_{(m,0,0)}\right) = \mathrm{E}\left(\min\left\{Q, V, \tilde{T}_1, \ldots, \tilde{T}_m\right\}\right) = \int_0^\infty e^{-(\lambda+\delta)r}\overline{F}_{\tilde{T}}(r)^{m}\,\mathrm{d}r,$$

$$m = k, \ldots, c-1. \tag{A.81}$$

We obtain (A.79)-(A.81) using the same logic as (A.4), (A.13), and (A.18). Note that equations for $\mathrm{E}\left(R_{(c+n,1,1)}\right)$; $\mathrm{E}\left(R_{(m,1,1)}\right)$, $m = k+n+1, ..., c+n-1$; and $\mathrm{E}\left(R_{(m,0,1)}\right)$, $m = k+n, ..., c+n-1$, are the same as (A.79)-(A.81) except that we remove the random variable V and the factor $e^{-\delta r}$ from the equations.

A.6 Notations

Table A.1: Frequently used notations listed in alphabetical order.

A	Set of absorbing states
A^c	Set of transient states
$\bar{b}_k$	Empirical k-partial busy period duration
B_k	k-partial busy period duraion
$\tilde{B}_{kk'}$	Residual k-partial busy period duration when there are $k' \geq k$ busy servers in the system
c	Scheduled number of servers (ambulances)
$c^{(\tau)}$	Scheduled number of servers (ambulances) in time segment τ
δ	Arrival rate of requested ambulances
e_n	Indicator whether the nth event is an arrival ($e_n = 1$) or a departure ($e_n = 0$)
$\mathrm{E}(X)$	The expected value of X
$F_X(x)$	CDF of the random variable X
$f_X(x)$	PDF of the random variable X
g_n	Indicator whether the requested ambulances are available right after the nth event
$\mathscr{G}_X(s)$	Probability-generating function of the random variable X
GH	Generalized hyperexponential distribution
GHE	Generalized hyper-Erlang distribution
H	Hyperexponential distribution
$H_{kk'}$	The expected number of lost calls during the residual k-partial busy period when there are currently $k' \geq k$ busy servers in the system
$\mathscr{L}_X(s)$	Laplace Transform of the PDF of X
$\mathscr{L}_{f(x)}(s)$	Laplace Transform of the function $f(X)$
λ	Arrival rate
$\lambda^{(\tau)}$	Arrival rate within time segment τ
L_k^c	Last event was a departure when there is currently k busy servers in the system
L_k	Last event was an arrival when there is currently k busy servers in the system
μ	Service rate
μ_k	Service rate in state k
μ_{new}	The new and increased service rate
$\mu_k^{(\tau)}$	State k service rate within time segment τ
μ_{early}	The service rate of prioritized ambulances
ν_n	Number of busy servers (ambulances) right after the nth event
ν_t	Number of busy servers (ambulances) at time t
N_k^c	Next event is a departure when there is currently k busy servers in the system
N_k	Next event is an arrival when there is currently k busy servers in the system
Ω	State space
Q	Call inter-arrival time
R_k	Sojourn time while there is currently k busy servers in the system
$\mathrm{SCV}(X)$	The squared coefficient of variation of X
T	Service time
$\tilde{T}$	Residual service time
$\mathrm{Var}(X)$	The variance of X
V	Time to the arrival of requested ambulances

APPENDIX B

Algorithms, Proofs, and Notations for QBD analysis

B.1 Algorithms

B.1.1 Erlang A Truncation Level

Algorithm B.1 shows the pseudo code of the algorithm by Ingolfsson and Tang (2012) that we use to compute the truncation level p for an Erlang A system such that the probability mass in the truncated upper tail is less than ϵ_h.

B.1.2 Erlang A Upper-tail Probability

Algorithm B.2 shows the pseudo code of the algorithm we use to compute the probability mass P at and above level $c + u_0$ for an Erlang A system where c is the number of servers and u_0 is an integer. This algorithm is inspired by Ingolfsson and Tang (2012) and calculates P with relative error tolerance ϵ_h.

B.2 Proofs for Section 3.8

B.2.1 Proposition 3.6

We complete the proof in four steps. In Step 1, we list properties of a well-studied class of square matrices called non-singular M-matrices. In Step 2, we show

AlgorithmB.1: Ingolfsson and Tang (2012) algorithm for Erlang A probabilities.

1. Input: $\lambda, \mu, \gamma, c, \epsilon_h$
2. Initialization: $p = \infty, \ell = 0, u = 0, q_{c-\ell} = 1, q_{c+u} = 1,$ converge = FALSE, $\Delta_2 = 1$
3. While not converge
4. If $q_{c+u} > q_{c-\ell}$ or $\ell = c$,
5. $u = u + 1$
6. $b = \lambda/(c\mu + u\gamma)$
7. $q_{c+u} = bq_{c+u}$
8. Normalize: $\Sigma = 1 + q_{c+u}$
9. $q_{c+u} = q_{c+u}/\Sigma, q_{c-\ell} = q_{c-\ell}/\Sigma$
10. If $b < 1$ Then $\Delta_2 = bq_{c+u}/(1-b)$
11. Else
12. $\ell = \ell + 1$
13. $a = (c - \ell + 1)\mu/\lambda$
14. $q_{c-\ell} = aq_{c-\ell}$
15. Normalize: $\Sigma = 1 + q_{c-\ell}$
16. $q_{c+u} = q_{c+u}/\Sigma, q_{c-\ell} = q_{c-\ell}/\Sigma$
17. End
17. If $\Delta_2 < \epsilon_h$ then converge = TRUE
18. Return $p = c + u$.

AlgorithmB.2: Upper-tail probability for an Erlang A system.

1. Input: λ, μ, γ, c, ϵ_h, ℓ_0, u_0
2. Initialization: $P = 0$, $\ell = 0$, $u = 0$, $q_{c-\ell} = 1$ $q_{c+u} = 1$ converge = FALSE; $\Delta_1 = 1$,
3 $\Delta_2 = 1$,
4. While not converge
5. If $q_{c+u} > q_{c-\ell}$ or $\ell = c$,
6. $u = u + 1$
7. $b = \lambda/(c\mu + u\gamma)$
8. $q_{c+u} = bq_{c+u}$
9. If $u \geq u_0$ Then $P = P + q_{c+u}$
10. Normalize: $\Sigma = 1 + q_{c+u}$
11. $q_{c+u} = q_{c+u}/\Sigma, q_{c-\ell} = q_{c-\ell}/\Sigma, P = P/\Sigma$
12. If $b < 1$ Then $\Delta_2 = \min(bq_{c+u}/(1-b), 1)$
13. Else
14. $\ell = \ell + 1$
15. $a = (c - \ell + 1)\mu/\lambda$
16. $q_{c-\ell} = aq_{c-\ell}$
17. Normalize: $\Sigma = 1 + q_{c-\ell}$
17. $q_{c+u} = q_{c+u}/\Sigma, q_{c-\ell} = q_{c-\ell}/\Sigma, P = P/\Sigma$
18. If $a < 1$ Then $\delta_1 = \min(aq_{c-\ell}/(1-a), 0)$
20. If $\ell = c$ Then $\delta_1 = 0$
21. End
22. If $(\Delta_1 + \Delta_2) < \epsilon_h/(1+\epsilon_h)$ and $u > u_0$ Then converge = TRUE
22. Return P.

that $M_\ell(\boldsymbol{I})$ is a non-singular M-matrix, for $\ell \in \mathbb{Z}^+$. In Step 3, we prove that $\mathbf{0} \leq \underline{\boldsymbol{R}}^{(\ell)} \leq \boldsymbol{R}^{(\ell)}$. In Step 4, we prove that $\boldsymbol{R}^{(\ell)} \leq \overline{\boldsymbol{R}}^{(\ell)}$.

Step 1.

Any square matrix E with non-positive real off-diagonal entries is an M-matrix if E can be expressed as $\mathrm{E} = s\boldsymbol{I} - \boldsymbol{F}$, where $\boldsymbol{F} \geq \mathbf{0}$ and $s \geq \rho(\boldsymbol{F})$, the spectral radius of $\boldsymbol{F}$ (Berman and Plemmons 1987, Page 133).

Lemma B.1. *Assume $\boldsymbol{X}$ is a square matrix with non-positive off-diagonal entries. Then:*

a) *$\boldsymbol{X}$ is a non-singular M-matrix if and only if there exists a vector $\boldsymbol{a} > \mathbf{0}$ such that $\boldsymbol{X}\boldsymbol{a} > \mathbf{0}$.*

b) *if $\boldsymbol{X}$ is a non-singular M-matrix, then $\boldsymbol{X}^{-1} \geq 0$.*

Proof: Follows from properties $I28$ and $N38$ in Berman and Plemmons (1987, Chapter 6).

Step 2.

From the structure of the infinitesimal generator matrix (3.1), $\boldsymbol{A}_0^{(\ell)} \geq \mathbf{0}$, $\boldsymbol{A}_2^{(\ell)} \geq \mathbf{0}$, and $\boldsymbol{A}_1^{(\ell)}$ is a matrix with non-positive diagonal and non-negative off-diagonal entries, for $\ell \in \mathbb{Z}^+$. Off-diagonal elements of matrix $M_\ell(\boldsymbol{I}) = -\boldsymbol{A}_1^{(\ell+1)} - \boldsymbol{A}_0^{(\ell+1)}$ are non-positive because all off-diagonal elements of $-\boldsymbol{A}_1^{(\ell+1)}$ and $-\boldsymbol{A}_0^{(\ell+1)}$ are non-positive. Graphical representations of the signs of the entries of $\boldsymbol{A}_0^{(\ell)}, \boldsymbol{A}_1^{(\ell)}$, and $M_\ell(\boldsymbol{I})$, for $\ell \in \mathbb{Z}^+$, are as follows (we use $+$ and $-$ signs to represent non-negative and non-positive entries, respectively):

$$\boldsymbol{A}_0^{(\ell)} = \begin{pmatrix} + & \dots & + \\ \vdots & \ddots & \vdots \\ + & \dots & + \end{pmatrix}, \boldsymbol{A}_1^{(\ell)} = \begin{pmatrix} - & \dots & + \\ \vdots & \ddots & \vdots \\ + & \dots & - \end{pmatrix}, M_\ell(\boldsymbol{I}) = \begin{pmatrix} + & \dots & - \\ \vdots & \ddots & \vdots \\ - & \dots & + \end{pmatrix}. \tag{B.1}$$

Row sums of the matrix $M_\ell(\boldsymbol{I})$ are positive because, for all $i \in \mathbb{Y}$ and $\ell \in \mathbb{Z}^{++}$:

$$\left(\boldsymbol{A}_1^{(\ell)}\right)_{i,i} = -\left(\sum_{j\in\mathbb{Y}, j\neq i}\left(\boldsymbol{A}_1^{(\ell)}\right)_{i,j} + \sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_0^{(\ell)}\right)_{i,j} + \sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_2^{(\ell)}\right)_{i,j}\right). \tag{B.2}$$

Therefore,

$$\sum_{j\in\mathbb{Y}}\left(M_\ell(\boldsymbol{I})\right)_{i,j} = -\sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_1^{(\ell+1)}\right)_{i,j} - \sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_0^{(\ell+1)}\right)_{i,j} = \sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_2^{(\ell+1)}\right)_{i,j} > \mathbf{0}. \tag{B.3}$$

The matrix $M_\ell(\boldsymbol{I})$ has non-positive off-diagonal entries and positive row sums ($\mathbf{a} = \mathbf{1}$ in Lemma B.1, Part a), so is a non-singular M-matrix.

Step 3.

For stochastic matrix $\boldsymbol{G}^{(\ell+1)}$, the row sums of $M_\ell(\boldsymbol{G}^{(\ell+1)})$ reduce to:

$$\sum_{j\in\mathbb{Y}}\left(M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)\right)_{i,j} = -\sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_1^{(\ell+1)}\right)_{i,j} - \sum_{j\in\mathbb{Y}}\sum_{k\in\mathbb{Y}}\left(\boldsymbol{A}_0^{(\ell+1)}\right)_{i,k}\left(\boldsymbol{G}^{(\ell+1)}\right)_{k,j} \tag{B.4}$$

$$= -\sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_1^{(\ell+1)}\right)_{i,j} - \sum_{k\in\mathbb{Y}}\left(\boldsymbol{A}_0^{(\ell+1)}\right)_{i,k}\sum_{j\in\mathbb{Y}}\left(\boldsymbol{G}^{(\ell+1)}\right)_{k,j} \tag{B.5}$$

$$= -\sum_{j\in\mathbb{Y}}\left(\boldsymbol{A}_1^{(\ell+1)}\right)_{i,j} - \sum_{k\in\mathbb{Y}}\left(\boldsymbol{A}_0^{(\ell+1)}\right)_{i,k} > 0, \quad \forall i \in \mathbb{Y}, \forall \ell \in \mathbb{Z}^+. \tag{B.6}$$

The last inequality holds as we showed in (B.3). As $\mathbf{0} \leq \underline{\boldsymbol{G}}^{(1)} \leq \boldsymbol{G}^{(1)}$ (inequality (3.23)) and $\boldsymbol{A}_0^{(\ell+1)} \geq \mathbf{0}$, then $M_\ell(\boldsymbol{G}^{(\ell+1)}) \leq M_\ell(\underline{\boldsymbol{G}}^{(\ell+1)})$, which implies that

$$0 < \sum_{j\in\mathbb{Y}}\left(M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)\right)_{i,j} \leq \sum_{j\in\mathbb{Y}}\left(M_\ell\left(\underline{\boldsymbol{G}}^{(\ell+1)}\right)\right)_{i,j}, \quad \forall i \in \mathbb{Y}, \forall \ell \in \mathbb{Z}^+. \tag{B.7}$$

Therefore, $M_\ell(\boldsymbol{G}^{(\ell+1)})$ and $M_\ell(\underline{\boldsymbol{G}}^{(\ell+1)})$ satisfy Lemma B.1, Part a) conditions and both are non-singular M-matrices with positive inverses:

$$M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)^{-1} > \mathbf{0} \text{ and } M_\ell\left(\underline{\boldsymbol{G}}^{(\ell+1)}\right)^{-1} > \mathbf{0}. \tag{B.8}$$

Define $\Delta M^{(\ell)} = M_\ell(\underline{\boldsymbol{G}}^{(\ell+1)}) - M_\ell(\boldsymbol{G}^{(\ell+1)}) \geq \mathbf{0}$. Using Miller (1981b) results, we obtain:

$$\left(M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right) + \Delta M^{(\ell)}\right)^{-1} = \tag{B.9}$$

$$M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)^{-1} - \frac{M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)^{-1} \Delta M^{(\ell)} M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)^{-1}}{1 + \mathrm{tr}\left(\Delta M^{(\ell)} M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)^{-1}\right)}, \forall \ell \in \mathbb{Z}^+, \tag{B.10}$$

where tr(.) is the sum of the diagonal elements of a matrix. Combining $M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)^{-1} \geq \mathbf{0}$ and $\Delta M^{(\ell)} \geq \mathbf{0}$ with (B.10) and (B.8), we conclude that $\mathbf{0} < M_\ell\left(\underline{\boldsymbol{G}}^{(\ell+1)}\right)^{-1} = \left(M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right) + \Delta M^{(\ell)}\right)^{-1} \leq M_\ell\left(\boldsymbol{G}^{(\ell+1)}\right)^{-1}$, and therefore $\mathbf{0} < \underline{\boldsymbol{R}}^{(\ell)} \leq \boldsymbol{R}^{(\ell)}$.

Step 4.

All off-diagonal entries of $M_\ell\left(\overline{\boldsymbol{G}}^{(\ell+1)}\right) + \boldsymbol{D}^{(\ell)}$ are non-positive, and all row sums of $M_\ell\left(\overline{\boldsymbol{G}}^{(\ell+1)}\right) + \boldsymbol{D}^{(\ell)}$ are positive (by the definition of $\boldsymbol{D}^{(\ell)}$, each row sum of $M_\ell\left(\overline{\boldsymbol{G}}^{(\ell+1)}\right) + \boldsymbol{D}^{(\ell)}$ is at least $\xi^{(\ell)}$), therefore $M_\ell\left(\overline{\boldsymbol{G}}^{(\ell+1)}\right) + \boldsymbol{D}^{(\ell)}$ is a non-singular M-matrix (Lemma B.1, Part a). As $\mathbf{0} \leq \boldsymbol{G}^{(\ell+1)} \leq \overline{\boldsymbol{G}}^{(\ell+1)}$ and $\boldsymbol{A}_0^{(\ell+1)} \geq \mathbf{0}$, then $M_\ell(\overline{\boldsymbol{G}}^{(\ell+1)}) \leq M_\ell(\boldsymbol{G}^{(\ell+1)})$, which implies that

$$\sum_{j\in\mathbb{Y}} \left(M_\ell(\overline{\boldsymbol{G}}^{(\ell+1)})\right)_{i,j} \leq \sum_{j\in\mathbb{Y}} \left(M_\ell(\boldsymbol{G}^{(\ell+1)})\right)_{i,j}, \quad \forall i \in \mathbb{Y}, \forall \ell \in \mathbb{Z}^+. \tag{B.11}$$

As shown in (B.6), $M_\ell(\boldsymbol{I})\mathbf{1} = M_\ell(\boldsymbol{G}^{(\ell+1)})\mathbf{1} > \mathbf{0}$. The definition of $\boldsymbol{D}^{(\ell)}$ guarantees that when we add $(\boldsymbol{D}^{(\ell)})_{i,i}$ to the left-hand side of (B.11), then we get a positive lower bound for $\sum_{j\in\mathbb{Y}} \left(M_\ell(\boldsymbol{G}^{(\ell+1)})\right)_{i,j}$. That is

$$0 < \sum_{j\in\mathbb{Y}} \left(M_\ell(\overline{\boldsymbol{G}}^{(\ell+1)})\right)_{i,j} + \left(\boldsymbol{D}^{(\ell)}\right)_{i,i} \leq \sum_{j\in\mathbb{Y}} \left(M_\ell(\boldsymbol{G}^{(\ell+1)})\right)_{i,j}, \quad \forall i \in \mathbb{Y}, \forall \ell \in \mathbb{Z}^+. \tag{B.12}$$

Therefore, $M_\ell\left(\overline{\boldsymbol{G}}^{(1\ell+)}\right) + \boldsymbol{D}^{(\ell)}$ is a non-singular M-matrix. Applying the Miller (1981b) formula we used in (B.10) for matrices $M_\ell\left(\overline{\boldsymbol{G}}^{(\ell+1)}\right) + \boldsymbol{D}^{(\ell)}$ and $\boldsymbol{G}^{(\ell+1)}$, one can show that $\boldsymbol{R}^{(\ell)} \leq \overline{\boldsymbol{R}}^{(\ell)}$.

B.2.2 Proposition 3.7

We prove this proposition in two steps. In Step 1, we prove that $\underline{\boldsymbol{x}}_0 > 0$ and $\boldsymbol{x}_0 \leq \overline{\boldsymbol{x}}_0$. In Step 2, we prove that $\underline{\boldsymbol{x}}_0 \leq \boldsymbol{x}_0$.

Step 1.

Assume we know $\boldsymbol{G}^{(1)}$. To solve (3.28), we define the following notations:

$$\boldsymbol{x}_{\text{cut}} = (x_{0,1}, ..., x_{0,p}), \boldsymbol{b}_{\text{cut}} = -(a_{p+1,1} + z_{p+1,1}, ..., a_{p+1,p} + z_{p+1,p}),$$

$$\boldsymbol{A}_{\text{cut}} = \begin{pmatrix} -a_{1,1} & \dots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{p,1} & \dots & -a_{p,p} \end{pmatrix}, \boldsymbol{Z}_{\text{cut}} = \begin{pmatrix} z_{1,1} & \dots & z_{1,p} \\ \vdots & \ddots & \vdots \\ z_{p,1} & \dots & z_{p,p} \end{pmatrix}.$$

All off-diagonal entries of $\boldsymbol{A}_{\text{cut}}$ are non-negative (by construction) and $\boldsymbol{Z} > \boldsymbol{0}$ (In QBDs of our context, all phases in Level 0 are directly accessible from Level 1, so $G^{(1)} > 0$, and there is at least one non-zero entry in each row of $\boldsymbol{A}_0^{(0)}$), therefore the off-diagonal entries of $\boldsymbol{K}_{\text{cut}} = \boldsymbol{A}_{\text{cut}} + \boldsymbol{Z}_{\text{cut}}$ are positive. We multiply both sides of (3.28) with -1 and obtain $-\boldsymbol{x}_{\text{cut}}\boldsymbol{K}_{\text{cut}} = -\boldsymbol{b}_{\text{cut}}$, which is a system of linear equations with a coefficient matrix, $-\boldsymbol{K}_{\text{cut}}$, that has negative off-diagonal entries and a right-hand-side vector, $-\boldsymbol{b}_{\text{cut}}$, that has positive entries. The vector $\boldsymbol{x}_{\text{cut}}$ is positive because (3.27) has a positive solution (Bright and Taylor 1995), and therefore $-\boldsymbol{K}_{\text{cut}}$ is a non-singular M-matrix as a result of Lemma B.1, Part a) in Section B.2.1.

Now, we prove that the solution of $\underline{\boldsymbol{x}}_0 \left(\boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0^{(0)}\underline{\boldsymbol{G}}^{(1)}\right) = \boldsymbol{0}$, $\underline{\boldsymbol{x}}_0$, is positive. We define $\underline{\boldsymbol{Z}} = \boldsymbol{A}_0^{(0)}\underline{\boldsymbol{G}}^{(1)}$ and write $\underline{\boldsymbol{x}}_0\underline{\boldsymbol{K}} = \boldsymbol{0}$ in the form of (3.28) by setting $\underline{x}_{0,p+1} = 1$ and removing the last equation. That is:

$$\left(\underline{x}_{0,1}, ..., \underline{x}_{0,p}\right)\left(\begin{pmatrix} -a_{1,1} & \dots & a_{1,p} \\ \vdots & \ddots & \vdots \\ a_{p,1} & \dots & -a_{p,p} \end{pmatrix} + \begin{pmatrix} \underline{z}_{1,1} & \dots & \underline{z}_{1,p} \\ \vdots & \ddots & \vdots \\ \underline{z}_{p,1} & \dots & \underline{z}_{p,p} \end{pmatrix}\right) =$$

$$-\left(a_{p+1,1} + \underline{z}_{p+1,1}, ..., a_{p+1,p} + \underline{z}_{p+1,p}\right). \tag{B.13}$$

Let:

$$\underline{\boldsymbol{x}}_{\text{cut}} = \left(\underline{x}_{0,1}, ..., \underline{x}_{0,p}\right), \underline{\boldsymbol{Z}}_{\text{cut}} = \begin{pmatrix} \underline{z}_{1,1} & \cdots & \underline{z}_{1,p} \\ \vdots & \ddots & \vdots \\ \underline{z}_{p,1} & \cdots & \underline{z}_{p,p} \end{pmatrix},$$

$$\underline{\boldsymbol{b}}_{\text{cut}} = -(a_{p+1,1} + \underline{z}_{p+1,1}, ..., a_{p+1,p} + \underline{z}_{p+1,p}).$$

The inequality $\mathbf{0} \leq \underline{\boldsymbol{Z}}_{\text{cut}} \leq \boldsymbol{Z}_{\text{cut}}$ holds, because $\boldsymbol{A}_0^{(0)} \geq \mathbf{0}$ and $\mathbf{0} \leq \underline{\boldsymbol{G}}^{(1)} \leq \boldsymbol{G}^{(1)}$ (inequality (3.23)). Therefore, $-\boldsymbol{K}_{\text{cut}} \leq -\underline{\boldsymbol{K}}_{\text{cut}}$, where $\underline{\boldsymbol{K}}_{\text{cut}} = \boldsymbol{A}_{\text{cut}} + \underline{\boldsymbol{Z}}_{\text{cut}}$. We multiply both sides of $-\boldsymbol{K}_{\text{cut}} \leq -\underline{\boldsymbol{K}}_{\text{cut}}$ with $\boldsymbol{x}_{\text{cut}} > \mathbf{0}$ and obtain $0 < -\boldsymbol{x}_{\text{cut}}\boldsymbol{K}_{\text{cut}} \leq -\boldsymbol{x}_{\text{cut}}\underline{\boldsymbol{K}}_{\text{cut}}$. The off-diagonal entries of $-\underline{\boldsymbol{K}}_{\text{cut}}$ are non-positive and $-\boldsymbol{x}_{\text{cut}}\underline{\boldsymbol{K}}_{\text{cut}} > \mathbf{0}$, and therefore the matrix $-\underline{\boldsymbol{K}}_{\text{cut}}$ is a non-singular M-matrix and its inverse is a positive matrix (Lemma B.1, Part b). The vector $-\underline{\boldsymbol{b}}_{\text{cut}}$ is also non-negative, therefore $\underline{\boldsymbol{x}}_{\text{cut}} = \underline{\boldsymbol{b}}_{\text{cut}}\underline{\boldsymbol{K}}_{\text{cut}}^{-1} \geq \mathbf{0}$. As the system is positive recurrent, so the probability of visiting each state is positive; that is $\underline{\boldsymbol{x}}_{\text{cut}} = \underline{\boldsymbol{b}}_{\text{cut}}\underline{\boldsymbol{K}}_{\text{cut}}^{-1} > \mathbf{0}$.

To prove $\underline{\boldsymbol{x}}_0 \leq \boldsymbol{x}_0$, we define $\Delta\boldsymbol{K} = \boldsymbol{K}_{\text{cut}} - \underline{\boldsymbol{K}}_{\text{cut}} \geq \mathbf{0}$. We know $\boldsymbol{x}_{\text{cut}} = \boldsymbol{b}_{\text{cut}}\boldsymbol{K}_{\text{cut}}^{-1}$ and $\underline{\boldsymbol{x}}_{\text{cut}} = \underline{\boldsymbol{b}}_{\text{cut}}(\boldsymbol{K}_{\text{cut}} - \Delta\boldsymbol{K})^{-1}$. As $-\boldsymbol{K}_{\text{cut}}^{-1} \geq \mathbf{0}$, $-(\boldsymbol{K}_{\text{cut}} - \Delta\boldsymbol{K})^{-1} \geq \mathbf{0}$, and $-\underline{\boldsymbol{b}}_{\text{cut}} \leq -\boldsymbol{b}_{\text{cut}}$, then $\underline{\boldsymbol{x}}_{\text{cut}} \leq \boldsymbol{x}_{\text{cut}}$ if we show that $-(\boldsymbol{K}_{\text{cut}} - \Delta\boldsymbol{K})^{-1} \leq -\boldsymbol{K}_{\text{cut}}^{-1}$. Using results from Miller (1981b), we obtain:

$$-(\boldsymbol{K}_{\text{cut}} - \Delta\boldsymbol{K})^{-1} = -\boldsymbol{K}_{\text{cut}}^{-1} + \frac{1}{1 + \text{tr}\left(-\Delta\boldsymbol{K}\boldsymbol{K}_{\text{cut}}^{-1}\right)}\boldsymbol{K}_{\text{cut}}^{-1}(-\Delta\boldsymbol{K})\boldsymbol{K}_{\text{cut}}^{-1}, \qquad \text{(B.14)}$$

As both $\Delta\boldsymbol{K}$ and $-\boldsymbol{K}_{\text{cut}}^{-1}$ are non-negative matrices, then

$$(1/1 + \text{tr}\left(-\Delta\boldsymbol{K}\boldsymbol{K}_{\text{cut}}^{-1}\right))\boldsymbol{K}_{\text{cut}}^{-1}(-\Delta\boldsymbol{K})\boldsymbol{K}_{\text{cut}}^{-1} \leq \mathbf{0}, \qquad \text{(B.15)}$$

so (B.14) results in $-(\boldsymbol{K}_{\text{cut}} - \Delta\boldsymbol{K})^{-1} \leq -\boldsymbol{K}_{\text{cut}}^{-1}$, and therefore $\underline{\boldsymbol{x}}_{\text{cut}} \leq \boldsymbol{x}_{\text{cut}}$. If we concatenate $x_{0,p+1} = 1$ to vectors $\underline{\boldsymbol{x}}_{\text{cut}}$ and $\boldsymbol{x}_{\text{cut}}$, we obtain $\underline{\boldsymbol{x}}_0 \leq \boldsymbol{x}_0$.

Step 2.

We proceed by showing that the solution of $\overline{\boldsymbol{x}}_0 \left(\boldsymbol{A}_1^{(0)} + \boldsymbol{A}_0^{(0)} \overline{\boldsymbol{G}}^{(1)} + \boldsymbol{H} \right) = \mathbf{0}$, $\overline{\boldsymbol{x}}_0$, is positive. Analogous to $\underline{\boldsymbol{K}}_{\text{cut}}$, we obtain $\overline{\boldsymbol{K}}_{\text{cut}}$ buy cutting the $(p+1)$th column and row of $\overline{\boldsymbol{K}}$. The definition of $\overline{\boldsymbol{K}}$ guarantees that $-\underline{\boldsymbol{x}}_0 \overline{\boldsymbol{K}}_{\text{cut}} > \mathbf{0}$. All off-diagonal entries of $-\overline{\boldsymbol{K}}_{\text{cut}}$ are non-positive and there is a positive vector, $\underline{\boldsymbol{x}}_0$, that satisfies Lemma B.1, Part a), therefore $-\overline{\boldsymbol{K}}_{\text{cut}}$ is a non-singular M-matrix, and therefore $-\overline{\boldsymbol{K}}_{\text{cut}}^{-1} \geq \mathbf{0}$ (Lemma B.1, Part b). The vector $-\overline{\boldsymbol{b}}_{\text{cut}}$ is also non-negative, therefore $\overline{\boldsymbol{x}}_0 = \boldsymbol{b}_{\text{cut}} \overline{\boldsymbol{K}}_{\text{cut}}^{-1} \geq \mathbf{0}$. As the system is positive recurrent, so the probability of visiting all states are positive; that is $\overline{\boldsymbol{x}}_0 = \boldsymbol{b}_{\text{cut}} \overline{\boldsymbol{K}}_{\text{cut}}^{-1} > \mathbf{0}$. In the same fashion as when we showed $-\underline{\boldsymbol{K}}_{\text{cut}}^{-1} \leq -\boldsymbol{K}_{\text{cut}}^{-1}$ in Step 1, one can confirm that $-\boldsymbol{K}_{\text{cut}}^{-1} \leq -\overline{\boldsymbol{K}}_{\text{cut}}^{-1}$. By construction, we know that $-\boldsymbol{b}_{\text{cut}} \leq -\overline{\boldsymbol{b}}_{\text{cut}}$, so $\boldsymbol{b}_{\text{cut}} \boldsymbol{K}_{\text{cut}}^{-1} \leq \overline{\boldsymbol{b}}_{\text{cut}} \overline{\boldsymbol{K}}_{\text{cut}}^{-1}$, or equivalently $\boldsymbol{x}_{\text{cut}} \leq \overline{\boldsymbol{x}}_{\text{cut}}$. If we concatenate $x_{0,p+1} = 1$ to the vectors $\overline{\boldsymbol{x}}_{\text{cut}}$ and $\boldsymbol{x}_{\text{cut}}$, we obtain $\boldsymbol{x}_0 \leq \overline{\boldsymbol{x}}_0$.

B.2.3 Proposition 3.8

This inequality holds for $\ell = 0$ as stated in Proposition 3.7. As induction hypothesis, assume that $\mathbf{0} \leq \underline{\boldsymbol{x}}_\ell \leq \boldsymbol{x}_\ell \leq \overline{\boldsymbol{x}}_\ell$ holds for $\ell = 1, ..., n$. We know $\mathbf{0} \leq \underline{\boldsymbol{R}}^{(n)} \leq \boldsymbol{R}^{(n)} \leq \overline{\boldsymbol{R}}^{(n)}$. We therefore find $\mathbf{0} \leq \underline{\boldsymbol{x}}_n \ \underline{\boldsymbol{R}}^{(n)} \leq \boldsymbol{x}_n \ \boldsymbol{R}^{(n)} \leq \overline{\boldsymbol{x}}_n \ \overline{\boldsymbol{R}}^{(n)}$, which results in $\mathbf{0} \leq \underline{\boldsymbol{x}}_{n+1} \leq \boldsymbol{x}_{n+1} \leq \overline{\boldsymbol{x}}_{n+1}$.

B.2.4 Proposition 3.9

As shown in Propositions 3.7 and 3.8, $\underline{\boldsymbol{x}}_\ell \leq \boldsymbol{x}_\ell \leq \overline{\boldsymbol{x}}_\ell$, $\forall \ell \in \mathbb{Z}^+$. Therefore, $\underline{\boldsymbol{x}}_\ell / \overline{c} \leq \boldsymbol{x}_\ell / c \leq \overline{\boldsymbol{x}}_\ell / \underline{c}$, or equivalently $\underline{\boldsymbol{\pi}}_\ell \leq \boldsymbol{\pi}_\ell \leq \overline{\boldsymbol{\pi}}_\ell$.

B.3 Notations

Table B.1: Frequently used notations listed in alphabetical order.

$\bar{.}$	Upper bound
$\underline{.}$	Lower bound
$(.)_{ij}$	Element ij
$\Delta.$	Difference between the upper bound and lower bound
$\mathbf{1}$	A column vector of ones of appropriate size
$\boldsymbol{A}_0^{(\ell)}$	Matrix of transition rates from phases in level ℓ to phases in level $\ell+1$
$\boldsymbol{A}_1^{(\ell)}$	Matrix of transition rates from phases in level ℓ to phases in level ℓ
$\boldsymbol{A}_2^{(\ell)}$	Matrix of transition rates from phases in level ℓ to phases in level $\ell-1$
c	Normalizing factor
ϵ_h	Desired maximum probability mass for the truncated upper tail of Class-1 Erlang A system
ϵ_ℓ	Desired maximum probability mass for the truncated upper tail of LDQBD
$\boldsymbol{G}^{(\ell)}$	$\boldsymbol{G}$ matrix for level ℓ
γ_1	Impatience rate of Class-1 customer
γ_2	Impatience rate of Class-1 customer
h	Phase variable
k	Truncation level
ℓ	Level variable
λ_1	Arrival rate of Class-1 customer
λ_2	Arrival rate of Class-2 customer
m	Parameter used to obtain estimates of the rate and $\boldsymbol{G}$ matrices
$M_\ell(.)$	A matrix function defined to calculate bounds on $\boldsymbol{R}^{(\ell)}$
μ_1	Service rate of Class-1 customer
μ_2	Service rate of Class-2 customer
$N_\ell(.)$	A matrix function defined to calculate bounds on $\boldsymbol{R}^{(\ell)}$
p	Maximum phase
π	Concatenation of probability vectors, $\{\pi_0, \pi_1, ...\}$
π_ℓ	Probability vector for level *ell*
$\boldsymbol{Q}$	The infinitesimal generator matrix
q_1	Number rate of Class-1 customer in the queue
q_2	Number rate of Class-2 customer in the queue
$\boldsymbol{R}^{(\ell)}$	Rate matrix for level ℓ
$\boldsymbol{R}_m^{(\ell)}$	Estimate of $\boldsymbol{R}^{(\ell)}$ given that the system does not visit levels $\ell+m$ and above
$\boldsymbol{R}_m^{(\ell)}$	Estimate of $\boldsymbol{G}^{(\ell)}$ given that the system does not visit levels $\ell+m$ and above
$\mathbb{S}$	The state space
s	Number of servers
s_1	Number of Class-1 customer receiving service
s_2	Number rate of Class-2 customer receiving service
$\boldsymbol{X}$	Matrix of unknowns
$\boldsymbol{x}$	Concatenation of before-normalization probability vectors, $\{\boldsymbol{x}_0, \boldsymbol{x}_1, ...\}$
$\boldsymbol{x}_\ell$	Probability vector for level *ell* before mormalization
$\mathbb{Y}$	Set of phases, $\{0, 1, ..., p\}$
$\mathbb{Z}^+$	Set of non-negative integers, $\{0, 1, ...\}$
$\mathbb{Z}^{++}$	Set of positive integers, $\{1, ...\}$