4-23-2018 9:30 AM

Statistical Applications in Healthcare Systems

Maryam Mojalal, *The University of Western Ontario*

Supervisor: David Stanford, *The University of Western Ontario*
Co-Supervisor: Richard Caron, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

Abstract

This thesis consists of three contributing manuscripts related to waiting times with possible applications in health care. The first manuscript is inspired by a practical problem related to decision making in an emergency department (ED). As short-run predictions of ED censuses are particularly important for efficient allocation and management of ED resources we model ED changes and present estimations for short term (hourly) ED censuses at each time point. We present a Markov-chain based algorithm to make census predictions in near future.

Considering the variation in arrival pattern and service requirements, we apply and compare three models which best describe our data. We provide hourly predictions up to 24 hours in a day which will provide suggestions to ED managers on how to prevent over-crowding in their system. We illustrate our approach using 22 months data obtained from the ED of a hospital in southwestern Ontario.

The next two manuscripts extend the theory underlying the Accumulating Priority queues (APQs). We focus on the queues with two classes of customers and Poisson arrivals. The first work in this topic derives the stationary waiting time distributions for the class of lowest-priority customers in an Affine Accumulating Priority queues (Affine APQs). APQs were first studied by Kleinrock (1964) and later revisited by Stanford et. al (2014) where they obtained explicit solution for the Laplace Stieltjes Transform (LST) of the stationary waiting times for all classes of customers.

All subsequent publications on APQs, have assumed that all arriving customers accumulate priority credits over time starting from the same initial value (assumed, without loss of generality, to be 0). Whereas, our model studies Affine APQs which assume different initial priorities (without loss of generality in a two-class setting we assume the lowest class starts with 0 credit and the higher class customers with positive credit a. In this work we determine the waiting time distributions for the lower class of customers with Poisson arrivals and general service and present some numerical results for special cases of $M/M/1$, $M/M/c$ and $M/D/1$. Inspired by health care applications, we have also considered a particular optimization problem related to the Affine APQ model, in order to select the optimum accumulation rate which allows for the lowest class customers to meet their associated KPIs.

We next focus on the Analysis of the Maximum priority processes in the context of Affine APQ. Maximum Priority Processes were first introduced in the context of APQs in Stanford et. al (2014). We derive the LST of the stationary steady state distributions of the Maximum Priority Processes as recursive functions and derive the explicit solutions for the LSTs in classical APQ (i.e. $a = 0$). We employ this argument to present a new approach to determine the LST of waiting time distribution for an APQ with two-classes of customers under the $M/M/1$ discipline. Since the Analysis of the Maximum Priority Processes in this work is done for the general class of Affine APQs, it has provided the grounds for future researches to obtain the LST of the waiting time distributions in the Affine APQs.

Keywords: Health care, Discrete time Markov chain, ED Census predictions, Regression with ARIMA errors, (Affine) Accumulating priority queue, Waiting time distributions, Optimization.

Co-Authorship Statement

1. Paper title: Discrete time Markov chain algorithm for short time predictions in an Emergency Department (Maryam Mojalal, Greg Zaric, David A. Stanford, Alim Pardhan)

I would like to acknowledge Dr. Greg Zaric who contributed to this chapter of my thesis by providing ideas. He guided the analysis and provided suggestions and feedback for improving the text. I would also like to acknowledge Dr. Stanford who brought insights and provided ideas to guide the analysis. Dr. Alim Pardhan assisted us in receiving data and provided feedback on the actual processes in an emergency department.

2. Paper title: The Lowest Priority Waiting Time Distribution in the Affine and the Delayed Accumulating Priority Queues (Maryam Mojalal, David A. Stanford, Peter Taylor, Richard J. Caron)

I would like to acknowledge Dr. Stanford who first identified the algorithm, and sketched out Theorem 4.3.3. Dr. Stanford also contributed with corrections to improve the text and recommendations to the content and relevance. Dr. Peter Taylor provided feedback on the accuracy of the content and methodology.

I would also like to acknowledge Dr. Caron, who brought up the idea of affine APQs for the first time, for providing feedback on the content and text.

3. Chapter title: The bivariate Maximum Priority Process in an Affine APQ (Maryam Mojalal, Na Li, David A. Stanford, Peter Taylor)

The idea of analysing bi-variate Maximum Priority Processes was proposed by Dr. Peter Taylor. His master's student Dan Daan initiated the work on the classical APQ and came up with an initial draft of the transition probabilities. Similarly, Dr. Na Li extended those initial equations for the Affine APQ. Dr. David Stanford contributed to this chapter with corrections to improve the text and recommendations to the content and relevance.

Acknowledgements

I would like to extend thanks to the many people who made this thesis possible.

First, I would like to convey my heartiest gratitude to the professors and staff of the Department of Statistics and Actuarial Sciences at Western University, who always welcomed me and my questions warmly; specifically, our wonderful Graduate Chair Dr. Kristina Sendova who inspired me in many ways.

Specific mention goes to my supervisors, Dr. Stanford and Dr. Caron who provided the opportunity of pursuing a PhD career for me, and without whose continuous guidance, support, and help my dissertation would not have been completed. I would like to thank Dr. Stanford for supporting me in attending conferences in Canada and Europe to present my work.

Very special thanks to Dr. Zaric for introducing me to the world of applied operations management; for being extremely encouraging, respectful, and approachable and for providing valuable guidance and insight towards my future career.

A special feeling of gratitude to my parents, my sister and my dearest brother, Mohammad, who have loved me and supported me in every step of my life. I am also thankful to my fellow graduate students in the department, with whom I shared thoughts and had many discussions.

Last but certainly not least, I wish to thank Dr. Matt Davison, Dr. Myron Hlynka, Dr. Hao Yu, and Dr. Amardeep Thind for agreeing to serve on my thesis committee and for reading my thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

ACF	Auto Correlation Function
APQ	Accumulating priority queue
ARIMA	Autoregressive Integrated Moving Average
c.d.f.	Cumulative distribution function
CTAS	Canadian Triage and Acuity Scale
DTMC	Discrete Time Markov Chain
ED	Emergency Department
FCFS	First come first serve
GLM	Generalized linear model
GS	Gaver-Stehfest numerical inversion algorithm
i.i.d.	Independent and identically distributed
KPI	Key Performance Indicators
LST	Laplace-Stieltjes Transform
MSE	Mean Square Error
PACF	Partial Autocorrelation Function
p.d.f.	Probability density function
r.v.	Random variable

Chapter 1

Introduction

The degree to which each individual is able to gain entry to a health care unit and to receive care and services would define the health care accessibility for each patient. While it is possible for some patients to wait for treatment, others may have their illness worsen during their wait, which ironically requires more complex and costly treatment in the future. Also, research has shown that overcrowding and long wait times in hospitals or emergency departments will decrease the quality of patients care (see for example [39], [51]). Therefore, it is necessary to plan strategies in advance or consider some interventions before or at the peak times to control and manage congestion.

Consequently, the ability to predict the time when an Emergency Department's (ED) overcrowding will occur, remains a high priority for many departments. A huge literature is available on the different methods (i.e time series, regression, statistical (machine) learning algorithms etc.) which have been used by researchers to accurately predict ED census in different time horizons. While long term ED census predictions are necessary for strategic and tactical planning purposes, short-term and even hourly predictions of the number of patients will help the managers or ED administrative to assist in capacity planning, meeting key performance indicators of the queue and plan ahead to consider possible interventions.

Meanwhile, other research has been conducted to study the effectiveness of some interventions to improve patient flow in emergency departments. A systematic literature review in this regard could be found in [58]. Classifying patients (customers) into different priority groups according to their urgencies for commencement of service has been one important action. Therefore, as a mathematical tool, a classical priority queue, which selects a customer of a lower priority class when no customers of higher priority classes are waiting, appear to be the most popular approach in such a system. However, this approach can cause the customers from lower classes to experience extremely long waiting times which may result in serious outcomes. As a result, the Canadian Triage and Acuity Scale (CTAS) [22] is applied to specify a time limit and a corresponding compliance probability for each class of patients.

As the need for having targets for service commencement of various customer classes in health care seems crucial, it is also important for queueing disciplines to factor both a patients urgency and incurred waiting time to choose the next customer to be served. Kleinrock's [28] "delay

dependent queue" in 1964 was the first method introduced for selecting customers for service based upon the maximum linearly accumulated amount of priority to that instant. While he had been able to derive the mean waiting times for each class of customers, actual waiting time distributions were elusive for 50 years until Stanford et al. [10] derived explicit results in 2014. They renamed these models as "Accumulating priority queues (APQ)" and obtained precise priority (for each class of customers) as service commences, after introducing "Maximum Priority Processes". The customers waiting time can be trivially recovered by scaling by the appropriate accumulation credit.

This thesis extends the research on the accumulating priority queues in two directions: The first direction is the introduction of Affine and Delayed variants of APQ models. In this approach, higher class customers will enter the system with an initial positive credit, a, therefore the lower priority classes spend some time before they start to compete with the higher class to enter the service. While considering the incurred wait times in a queue under APQ discipline reduces the waiting times on average for lower class patients as compared with the classical priority queues, "Affine APQ's" would be a middle ground. We also prove that at least for the two-class case, a full equivalence exists between the affine APQ and what we call the "Delayed APQ". In the delayed APQ approach, both classes start with 0 credit upon arrival but customers from the lower priority class spend some time before they start to accumulate (positive) priority credit. Such situations arise (in the health care context) when it is felt that lower-priority patients' waiting times are only of concern once they reach a threshold, and hence only should earn priority from that point onward.

In this thesis, we present an algorithm to obtain the waiting time distributions for the lower priority class customers in the Affine APQ setting under Poisson arrival and General common service time distributions. We present some numerical examples and illustrate these waiting time distributions in graphs as comparison with the classical APQ and classical priority queue disciplines. Furthermore, we have seen the application of this algorithm in the context of an optimization example to select the optimum accumulation rate for different delay time and occupancy levels.

Deriving the waiting time distributions for the higher priority class in the Affine APQ setting is still an open problem. Therefore, the second direction in which we build upon existing theory, pertains to how to define and analyse the maximum priority processes in an Affine APQ. Since obtaining the exact priority of customers at the time of entrance into the service is the key element in deriving the waiting times both higher and lower priority classes, we aimed at deriving the LST of the stationary distributions for all possible sets of states in an Affine APQ under M/M/1 discipline. We were able to write them in a set of recursive equations for the Affine APQ. We could also demonstrate as an application of this analysis how the LST functions for the special case when the initial credit is 0 (which reduces an Affine APQ to an APQ) can be derived. We linked these results to obtain the waiting time distribution for classical APQ. This introduces a new approach in studying the APQs and performs as an alternative. However, it can potentially be expanded to account for the general class of Affine APQs.

1.1 Outline of this thesis

A detailed review of related literature is presented in Chapter 2, including a short introduction to time series, Poisson regression and queueing theory, theory and some background information on accumulating priority queues, introduction to the Affine APQs, and a discussion of the Gaver-Stehfest numerical inversion algorithm.

In Chapter 3, we present a Markov-chain based algorithm to make near term census predictions for an ED. Considering the variation in arrival pattern and service requirements in an ED, we apply and compare three models which best describe our data. We provide hourly predictions up to 24 hours in a day which will provide suggestions to ED managers on how to prevent over-crowding in their system. We illustrate our approach using 22 months data obtained from the ED of Hamilton hospital.

Chapter 4 presents detailed background on Affine APQs and introduces the concept of delayed APQ. With an initial discussion on the waiting times for the delayed APQ, the linkage between the Affine APQs and delayed APQs is shown in a theorem. An algorithm for the derivation of stationary waiting time distributions in general service (one server) and Poisson arrival is introduced. This algorithm has also been expanded to include multiple Exponential servers. At the end, inspired from health care applications, a particular optimization problem related to the model, namely, the selection of the optimum accumulation rate which allows for the lowest class customers to meet their KPIs, is considered.

In chapter 5, we study the maximum priority processes in an Affine APQ. We derive a set of recursive equations for the LST stationary distributions of the states in an Affine APQ. We obtain an explicit solution for a special case when the initial class one credit a is set to be 0 (i.e. the classical APQ).

The main contributions are summarized in Chapter 6, as well as some future research directions.

Chapter 2

Preliminaries

In this chapter a review is given on the underlying methods that are required for the development of the presented later chapters. Since we develop models related to both statistical and queueing theory methodologies, this chapter is divided into two main sections: "Statistical models", and "Markov chains and queueing theory". In each section we focus on the main results and frequently used concepts in later chapters.

2.1 Statistical Models

We start this section with a brief introduction to time series models, specifically Autoregressive Integrated Moving Average (ARIMA) models, and Poisson regression. Poisson regression is one of the important models in the class of generalised linear models. These models have been frequently used in short and long term predictions of stochastic phenomena in stock market, traffic flow, weather forecasts, health care management, etc. We will follow with a brief review of the regression models with ARIMA errors which will be used in the next chapter.

2.1.1 Time series models

In this part, we'll describe some important features that we have considered when describing and modeling a time series in the following chapters. Time series are the data type which arise when a process is measured repeatedly and at equal or near equal time intervals. It has many research applications in health care [16] and other scientific areas. One of the earliest recorded series is the monthly sunspot numbers studied by Schuster in 1906 [6].

The time series techniques which are commonly used in health care analysis are moving average models, such as ARIMA, and smoothing techniques. For instance, the Box - Jenkins

ARIMA model (introduced by Box and Jenkins in 1970) which is commonly used in fitting forecasting models when dealing with a non-stationary time series, has been used extensively in health forecasting [37]. A time series is called "stationary" if its statistical properties, such as its mean, variance and autocorrelations, remain constant in time; otherwise it would be called "non-stationary". Various important concepts which we will be using frequently in this thesis are listed as follows:

White noise A particularly useful white noise series, w_t, is the Gaussian white noise wherein the w_t are independent normal random variables with mean 0 and variance σ_w^2. This can be denoted by

$$w_t \overset{i.i.d}{\sim} N(0, \sigma_w^2). \tag{2.1}$$

The auto covariance function Let x_t denote the value of a time series at time t. The auto covariance function is the second moment product [59]

$$\gamma_x(s,t) = cov(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)] \tag{2.2}$$

for all s and t where $E(x_s) = \mu_s$ and $E(x_t) = \mu_t$.

The auto correlation function (ACF) The ACF function is a normalized version of the auto covariance function which gives correlations between x_t and $x_s = x_{t-h}$ for $h = 1, 2, ..., t-1$ and is defined as

$$\rho(s,t) = \frac{\gamma_x(s,t)}{\sqrt{\gamma_x(s,s)\gamma_x(t,t)}}. \tag{2.3}$$

The ACF can be used to identify the possible structure of time series data, as it measures the linear predictability of the series at time t, say x_t, from the value of the series at time s. The ACF of the residuals for a model is also useful to check. The ideal for a sample ACF of residuals is that there are no significant correlations for any lag.

Partial Autocorrelation Function (PACF) If x_t is a Gaussian process, the partial autocorrelation function between x_t and x_{t-h} is defined as the correlation between x_t and x_{t-h}, conditional on $x_{t-h+1}, ..., x_{t-1}$, the set of observations that come between the time points t and $t-h$.

A (weakly) stationary time series [59] A weakly stationary time series, x_t, is a finite variance process such that
(i) the mean value function, μ_t, is constant and does not depend on time t, and
(ii) the auto covariance function, $\gamma(s,t)$, defined in (2.3) depends on s and t only through their difference $|s-t|$.

The Autoregressive Model (AR) The fundamental idea in these models is that x_t, the current series value, can be modeled as a function of the p previous lags. When p specifies the order of the model. It is abbreviated as $AR(p)$ and is written as,

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t, \tag{2.4}$$

where w_t is the white noise, and ϕ_i are fixed parameters of the model. The back shift operator is defined as $B^h x_t = x_{t-h}$. If we let $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$ be the autoregressive operator, then Equation (2.4) is also written as:

$$\phi(B)x_t = w_t. \tag{2.5}$$

The Moving Average model (MA) The qth order moving average model, abbreviated as $MA(q)$ is written as,

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}, \tag{2.6}$$

where w_t is the white noise and $\theta_i, i = 1, 2, ..., q$ are parameters of the model. If we define $\theta(B)$ as the moving average operator then Equation (2.6) could be written as:

$$x_t = \theta(B)w_t. \tag{2.7}$$

The ARMA model A time series is ARMA of order p and q abbreviated as ARMA(p, q), if it is stationary and

$$\phi(B)x_t = \theta(B)w_t, \tag{2.8}$$

where $\phi_p \neq 0$, $\theta_q \neq 0$.

A non-seasonal ARIMA model, abbreviated as $ARIMA(p, q)$ combines an $AR(P)$ and a $MA(q)$ processes but only after integrating them to transform the process into a stationary process.

The seasonal ARIMA model The seasonal ARIMA model incorporates both non-seasonal and seasonal factors in a multiplicative model. In this model another set of autoregressive, integration and moving average parameters are incorporated to consider the seasonality of process. The shorthand notation for the model is $ARIMA(p, d, q) \times (P, D, Q)^S$, with p, d and q as the non-seasonal AR, differencing and MA orders, while P, D and Q are the seasonal orders respectively. The model could be written more formally as:

$$\Phi(B^S)\phi(B)(1 - B)^d(1 - b^S)^D x_t = \Theta(B^S)\theta(B)w_t, \tag{2.9}$$

where $\Phi(B^S) = 1 - \Phi_1 B^S - ... - \Phi_P B^{PS}$ and $\Theta(B^S) = 1 + \Theta_1 B^S + ... + \Theta_Q B^{QS}$ are seasonal AR and MA operators.

2.1.2 Poisson regression model

Generalized linear models (GLM), popularized by McCullagh and Nelder in 1982, are extensions of linear models which can be used not only to model the data coming from normal distributions, but also non-normally distributed random variables. If the objective of a simple linear model is to model the expected value of a continuous Normally distributed variable, Y, as a linear function of the continuous (fixed) predictor, X, $E(Y_i) = \beta_0 + \beta_1 x_i$, in GLM models, the response variable Y_i is assumed to follow an exponential family distribution with mean μ_i, which is assumed to be some (often nonlinear) function of $x_i^T \beta$.

In this manner, GLM models can be used when the response variable, Y_i , takes any type of values (e.g., continuous, binary, count) and the predictors are connected to the response via a function called the "link function". The link function specifies how the expected value of the response relates to the linear predictor of explanatory variables; e.g., $\eta = g(E(Y_i)) = E(Y_i)$ for linear regression, or $\eta = \log(E(Y_i))$ for Poisson regression.

In Poisson regression the response variable is count and assumed to have a Poisson distribution that is $Y_i \sim Poisson(\mu_i)$ for $i = 1, ..., n$ and $E(Y_i) = \mu$. Therefore the model is written as

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = x^\top \beta, \tag{2.10}$$

where $X = (X_1, X_2, ..., X_k)^\top$ are explanatory variables.

2.1.3 Regression with ARIMA errors model

One of the main assumptions in an ordinary linear regression model is the assumption of uncorrelated error terms (residuals). However, it is possible that the errors of a regression model have a time series structure or have high autocorrelation which requires some modifications to our model and our way of analysis.

In this procedure the underlying model is called a "Regression model with auto correlated errors" [59]. That is, consider the model as:

$$\mathbf{y} = X\beta + \mathbf{e}, \tag{2.11}$$

where, $\mathbf{y}$ is an $n \times 1$ vector, X is the $n \times r$ regression covariate matrix (fixed input), β is an $r \times 1$ vector of regression parameters and n is the number of observations. The $n \times 1$ error vector, $\mathbf{e}$, is a process with some covariance implying that the error terms are correlated (i.e. are not white noise, w_t).

Now, if we have a pure $AR(p)$ error we can write:

$$e_t = \Phi^{-1}(B) w_t, \qquad t = 1, ..., n \tag{2.12}$$

where $\Phi(B)$ is the linear transformation that, when applied to the error process, produces the white noise w_t. Now the original model could be written as :

$$y_t^* = \mathbf{x}_t^* \beta + w_t \tag{2.13}$$

where $y_t^* = \Phi(B) y_t$ and $\mathbf{x}_t^* = \Phi(B)\mathbf{x}_t$. Once this transformation is done on the original data, an ordinary least squares regression model could be fitted to that to derive estimations for the parameters of the model.

$$\hat{\beta} = (X^{*\top} X^*)^{-1} X^{*\top} \mathbf{y}^*, \tag{2.14}$$

where $X^* = [\mathbf{x}_1^*, \mathbf{x}_2^*, ..., \mathbf{x}_n^*]^\top$ and $\mathbf{y}^* = [y_1^*, y_2^*, ..., y_n^*]^\top$.

2.2 Queueing theory

We start this section with a brief introduction to queueing systems and the mathematical study of these systems which is called queueing theory. We will focus on the study of a particular kind of queueing systems, namely, priority queueing systems, where these systems are useful particularly for the situations when certain kind of customers should be given faster access times to the server(s). We will follow with a review of the accumulating priority queue (APQ) as in Stanford et al. [10] which was first introduced as the time-dependent priority queues in Kleinrock [29]. We will review the single server APQ in Stanford et al., the homogeneous multi-server APQ in Sharif et al. [8], the Preemptive APQs in Fajardo and Drekic [5] and the nonlinear APQ in Li et al. [36]. Finally, as the rest of this thesis studies the Affine APQs, we will give an introduction to Affine APQs.

Furthermore, there will be a brief introduction to the Laplace-Stieltjes transform and a review of the Gaver-Stehfest numerical inversion algorithm as in [20, 40]. This Algorithm is used in this thesis many times to numerically invert the Laplace-Stieltjes transform (LST) of the waiting times' distribution functions.

2.3 The mathematical study of queueing systems

In order to completely develop a mathematical queueing model, we must identify those two fundamental processes that describe the arrival as well as the discipline of the service which fulfills the service requirements of the customers. Queueing theories are mainly concerned with the study of the systems that are limited in resources, where the probability of congestion is great.

The arrival process is generally described as the probability distribution of the inter-arrival times of the customers, which is denoted by $A(t)$. Service times however, are denoted by $B(t)$ which specify the service requirements of the customers. By using a notation which was first introduced by Kendall (1951), a queueing system is labelled as $A/B/m/c$, where letters A and B specify the arrival and service distributions. Conventionally, M stands for exponential distribution and G for unspecified "General" distribution. Furthermore, m specifies the number of servers in the system and c is the capacity of the queueing system, when c is not explicitly specified, it is presumed there are infinite waiting room capacity.

In this thesis, the inter-arrival times are assumed to be independent, service durations are independent, and the service durations are independent of the inter-arrival times. Other standard

assumptions considered in this thesis are as follows:

First and foremost is that all of our models describe Non-Preemptive work-conserving queues. Under this discipline, the work requirements of the customers are unaltered by the passage of time and the server never idles as long as there is work to be done. The non-preemptive discipline is such that a customer entering service will not be interrupted until service completion. Customers don't balk (i.e. decide not to enter a queue upon arrival if it is too long to suit them) or renege (i.e. decide to leave the queue after losing patience) which means no work (service requirement) is created or destroyed within the system. The simplest case to consider is the first-come-first-served (FCFS) service discipline. Service discipline is another very important characteristic of a queueing system, which governs the order of service of the customers.

The second assumption is that all models are operating in a stable regime; that is, the long-run service capacity exceeds the long-run demand. Also, we assume that the queues have operated sufficiently long to have reached a stationary steady state.

An important note to keep in mind is that due to the nature of queues in health care settings, systems are operating close to 100% utilization, which may not be a stable regime. Furthermore, balking and reneging of patients or change of their priorities due to health status may occur. Therefore, some of the results gained under standard assumptions may not apply to all systems, and it is up to the decision maker to take these into account based on the numerical results of any analysis.

Finally, before presenting some key distributional results within the $M/G/1$ framework, I briefly review Laplace-Stieltjes transforms (LSTs) and Gaver-Stehfest algorithm here. LSTs are widely used in probability and in the following chapters of this thesis.

Laplace-Stieltjes transforms

The Laplace-Stieltjes transforms (LST), often simply called Laplace transforms, are important tools when dealing with the distribution function of a nonnegative random variables and are extensively used in the following chapters. Therefore, given that F(x) is a distribution function defined by $F(x) = P(X \leq x)$, the corresponding LST is

$$\tilde{f}(s) = E(e^{-sX}) = \int_0^{+\infty} e^{-sx} dF(x) \tag{2.15}$$

for all s for which this integral converges. Also there is a one to one correspondence between a distribution function and its associated Laplace transform, such that it is often possible to invert a Laplace transform to recover its corresponding distribution functions both analytically by means of tables of inversion formulas or numerically by means of numerical algorithms such as the Gaver-Stehfest algorithm which has been used in this thesis and will be introduced shortly.
From equation (2.15), it can be concluded that if $\tilde{f}(s)$ is n times differentiable at the origin, then $E(X^n) = (-1)^n \tilde{f}^{(n)}(s)|_{s=0}$.

LSTs frequently appear in waiting time analysis in queueing systems. One reason according to Kleinrock [29] is that they arise naturally in the solution and the second reason is that they greatly simplify the calculations and oftentimes they are the only tools we have available for proceeding with a solution at all.

Gaver-Stehfest Algorithm of Inverse Laplace Transforms

Numerical inversion of Laplace transforms is crucial for many applications. The Gaver-Stehfest algorithm which was initially proposed by Gaver [40] and refined by Stehfest [20] is one of the most powerful algorithms for this purpose [63]. The Gaver-Stehfest method uses the summation

$$f(t) \approx \frac{\ln(2)}{t} \sum_{n=1}^{L} K_n \tilde{f}(\frac{n \ln(2)}{t}),$$

where $\tilde{f}(\cdot)$ is the Laplace transform of $f(t); t \geq 0$, a real valued function. The coefficient K_n depends only on the (necessarily even) number of expansion terms, $L \in \mathbb{N}$, and is given by

$$K_n = (-1)^{n+\frac{L}{2}} \sum_{k=\lfloor \frac{n+1}{2} \rfloor}^{\min(n, \frac{L}{2})} \frac{k^{L/2}(2k)!}{(L/2-k)!k!(k-1)!(n-k)!(2k-n)!},$$

when $\lfloor \cdot \rfloor$ is the floor function. These coefficients, as derived by Gaver, are combinatorial terms arising in order statistics, with the interesting by-product that they always sum to zero. In fact for each L, half of K_n's are positive and the other half are negative.

Typically $L = 8$ points provide two significant digits of accuracy, which is quite adequate for assessing waiting times.

In the context of waiting time distributions where $W_n(t)$ is the distribution function and $w_n(t)$ is the probability mass function and, in light of $\tilde{W}_n(s) = \frac{\tilde{w}_n(s)}{s}$ as a standard property of Laplace transforms, we have

$$W_n(t) \approx \frac{\ln(2)}{t} \sum_{n=1}^{L} K_n \frac{\tilde{w}_n(\frac{n \ln(2)}{t})}{(\frac{n \ln(2)}{t})} = \sum_{n=1}^{L} \frac{K_n}{n} \tilde{w}_n(\frac{n \ln(2)}{t}), \tag{2.16}$$

If $L = 6$ the coefficients are, 1, −49, 366, −858, 810 and −270. If $L = 8$ the coefficients are −1/3, 145/3, −906, 16394/3, −43130/3, 18730, −35840/3 and 8960/3.

2.3.1 Some of the fundamental results for $M/G/1$ queues

The $M/G/1$ queueing system is characterized by a Poisson arrival process at a mean rate of λ arrivals per unit of time and a "general" service time, G, distribution, $B(x)$. If q_n is the number of customers left behind by the departure of customer C_n, and v_n is the number of customers who enter during this customer's service time, x_n, the relationship among random variables for a FCFS $M/G/1$ system is formulated as

$$q_{n+1} = \begin{cases} q_n - 1 + v_{n+1} & \text{if } q_n > 0, \\ v_{n+1} & \text{if } q_n = 0. \end{cases} \tag{2.17}$$

The distribution function (c.d.f) of the inter-arrivals is given by

$$F(t) = 1 - e^{-\lambda t}, \qquad t \geq 0. \tag{2.18}$$

Furthermore, we let X represent the generally distributed service time random variable, then the corresponding c.d.f and LST would be denoted by

$$B(x) = P(X \leq x) \qquad \text{and} \qquad \tilde{B}(s) = E(e^{-sX})$$

respectively.

Therefore, if we let $p_{ij} = P\{q_{n+1} = j | q_n = i\}$ denote the one step transition probabilities observed only at departure instants, the matrix of transition probabilities takes the following form:

$$P = \begin{bmatrix} k_0 & k_1 & k_2 & k_3 \dots \\ k_0 & k_1 & k_2 & k_3 \dots \\ 0 & k_0 & k_1 & k_2 \dots \\ 0 & 0 & k_0 & k_1 \dots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix} \tag{2.19}$$

where k_i is the probability of i arrivals during the service time x and is obtained as

$$k_i = P\{v_{n+1} = i\} = \int_0^\infty \frac{(\lambda x)^i}{i!} e^{-\lambda x} b(x) dx.$$

Steady state probability vector $\pi = \{\pi_0, \pi_1, ...\}$ can be found as (see Equation 5.16 in [14])

$$\pi_i = \pi_0 k_i + \sum_{j=1}^{i+1} \pi_j k_{i-j+1}; \qquad i = 0, 1, 2, \dots . \tag{2.20}$$

Now, let T denote the duration of a typical busy period (i.e. the unbroken period when the backlog of work is greater than zero). By considering the idea of an initial queue and pseudo busy periods as in Conway et al. [61] (Section 8-3), the LST of T, $\tilde{\Gamma}(s) = E(e^{-st})$, is the solution to the functional equation

$$\tilde{\Gamma}(s) = \tilde{B}(s + \lambda(1 - \tilde{\Gamma}(s))). \tag{2.21}$$

We employed the fact that the distribution of the busy period length does not depend on the selection discipline unless the discipline entails extra processing, or insert idleness. Hence, moments of T can be obtained after differentiation of the LST in Equation (2.21):

$$E(T) = \frac{E(X)}{1 - \rho} \qquad \text{and,} \qquad E(T^2) = \frac{E(X^2)}{(1 - \rho)^3}, \tag{2.22}$$

where $\rho = \lambda E(X)$ is known as the traffic intensity and for values of $\rho < 1$, the queue is *stable* or *stationary* and the busy period had finite lengths with probability one (e.g., see Takács (1962, Theorem 3, p. 58)). For a stationary queueing system, ρ could be interpreted as the long-run fraction of time that the server is busy.

Although the above results are essential to this thesis, sometimes it will also be useful to consider a more general kind of busy period which is initiated by a processing time other than the subsequent service times.

In particular, we refer to *delay busy* periods, where the initiating task is called a delay and the time spent processing jobs is a delay busy period. Therefore, let T_d represent the duration of a delay busy period. Furthermore, let X_0, the initial delay, have density function $B_0(x)$ and LST function $\tilde{B}_0(s)$. Then, the LST of T_d would be

$$\tilde{\Gamma}_0(s) = \tilde{B}_0(s + \lambda(1 - \tilde{\Gamma}(s))), \tag{2.23}$$

where $\tilde{\Gamma}(s)$ is the solution to Eq (2.21). If $\rho < 1$ the limiting distributions of certain random variables such as the waiting time distribution of the n-th arriving customer, W_n, are known to exist (see Takács, 1962, Theorem 10, p. 69).

The associated LST is given by the *Pollaczek-Khinchin formula* for the $M/G/1$ system as:

$$\lim_{n\to\infty} \tilde{W}_n(s) = \tilde{W}(s) = \frac{s(1 - \rho)}{s - \lambda + \lambda\tilde{B}(s)}. \tag{2.24}$$

Also if $\tilde{W}_{BP}(s)$ is the waiting time LST for the customers arriving during the busy period, then an alternative for the above representation could be

$$\tilde{W}(s) = (1 - \rho) + \rho\tilde{W}_{BP}(s), \tag{2.25}$$

where it immediately follows that

$$\tilde{W}_{BP}(s) = \frac{(1-\rho)(1-\tilde{B}(s))}{E(X)(s-\lambda(1-\tilde{B}(s)))}.$$

2.3.2 An introduction to priority queueing systems

A queueing discipline which chooses the order of service based on some function of group membership is usually referred as a priority queueing discipline. In such setting, we assume that the arriving customers belong to one of distinct priority classes indexed by i, $(1, 2, ..., N)$, where the larger value of the index is associated with the higher priority class.

The aim of imposing a priority structure on the customer arriving is to provide treatment to the higher priority class at the expense of the lower classes. In the health care setting, the less tolerant patients (i.e., with more severe conditions) are assigned to higher priority class and should be treated sooner.

Throughout this thesis, we use the symbol C_i which is to be read as "class i customer" and we say C_k's are prioritized over C_j's whenever $k < j$.

A *static* priority queue (*classical priority* queue) is a priority queueing system where the priority function is fixed for a specific class and is static in time. Whereas, in a linear *time-dependent priority* system the priority function, in its general form (as in Hsu 1970 [25]), is defined as

$$q_i(t) = a_i + b_i(t - \tau_i), \qquad t \geq \tau_i, \qquad i = 1, 2, ..., N, \tag{2.26}$$

where the argument t represents time, and τ_i is the arrival time of a C_i. The set of class-dependent constants $\{a_i\}_{i=1}^{N}$ are arranged so that $a_1 > a_2 > ... > a_N$. These types of priority queues which are dependent on t have frequently been termed in the literature as *dynamic priority* discipline.

The first static priority queue was studied by Cobham in 1954 [3] and was later rigorously analyzed by many other queueing theorists. For a detailed analysis of this type of queues we refer the interested reader to the texts by Conway et al. [61] and Kleinrock [29]. In addition, another important feature of priority queues is based on the decision of whether or not to interrupt the customer in service for another higher priority customer arriving to the system. This kind of queueing systems are called *Preemptive queues* which have been studied in detail along with a thorough literature review in [4].

The first implementation of a dynamic priority queue was done by Jackson [53], [54], [55]. In his articles, Jackson considered a discrete-time queueing system and derived bounds for the mean waiting time of a class k customer. Later in 1962, he obtained an approximation for the waiting time distributions.

Kleinrock (1964) was the first who studied a dynamic priority discipline under a continuous-time frame-work. He studied linearly time-dependent queues with a set of positive variable

parameters (slope), b_i such that $b_1 \geq b_2 \geq ... \geq b_N \geq 0$, which are at the disposal of the queue administrative and allow them to adjust the relative waiting times of each priority group.

Hsu in 1970 [25] continues Kleinrock's work by deriving a result for a case when a unit's priority is decreased from zero linearly with time in proportion to a negative rate assigned to the unit's class (i.e $0 \geq b_1 \geq b_2 \geq ... \geq b_N$).

In 1967 Kleinrock and Finkelstein [30] extended their initial work by considering non-linear power-law functions of the form

$$q_i^r(t) = b_i\,(t - \tau_i)^r, \qquad t \geq \tau_i$$

when $r \geq 0$. They were able to obtain the expected value of the waiting time in this r-th order non-linear time dependent setting.

In the modern literature, Kleinrock's time-dependent priority queue has been rephrased as "Accumulating priority queues" initially by Stanford et al. (2015) when they revisited this problem after about three decades and studied this problem under more general assumptions.

2.4 Accumulating Priority Queue (APQ)

According to Stanford et al. [10], the specification of the APQ, key assumptions and results under a $M/G/c$ discipline are described as follows.

Assume there are $N \geq 2$ classes of customers and one or $c \geq 2$ servers in the system. Customers of class i arrive independently at the queue as a Poisson process with rate $\lambda_i, i = 1, 2, ..., N$. The accumulation function is

$$q_i(t) = b_i\,(t - \tau_i), \qquad t \geq \tau_i \tag{2.27}$$

which means upon arrival at τ_i, a customer of class i starts accumulating priority at rate b_i, $(b_1 > b_2 > ... > b_N > 0)$. Their accumulated priority, therefore, would be according to $q_i(t)$ and, when a server is available, the next customer to be served is the one with the highest priority at that instant. This is a non-preemptive system.

This problem was addressed and solved after defining the *Maximum Priority Process* (see [10], Definition 3.1). If we let $c = 2$, the bi-variate Maximum Priority Process $M(t) = (M_1(t), M_2(t)), t \geq 0$ in a two class setting is defined as an upper bound for the priorities of the queued customers from each class, given only the knowledge of arrival times and their accumulated priorities at those times.

The key undergoing idea to initiate the analysis is to find out the connection between the Poisson arrival processes and the accumulated priorities of the customers still waiting for service from either classes. Since the arrival process is Poisson, the accumulated priorities will be

distributed as independent Poisson processes with rate λ_i/b_i for class i; $i = 1, 2$ on the intervals $[0, M_i(t)]$ where $M_i(t)$;$i = 1, 2$ is the Maximum Priority Process.

As a result a waiting customer with a priority on interval $[0, M_2(t)]$ will be of class one with probability $\frac{\lambda_1(b_2/b_1)}{(\lambda_1(b_2/b_1)+\lambda_2)}$ independently of the class of all other customers present in the queue.

This important point leads to the next interesting result which is the rate at which a class one customer overtakes all waiting class two customers in the queue. This is also called the accreditation rate and that class one customer is called an accredited class-1 customer. Furthermore, the accreditation interval consists of the service time of a non-accredited customer (could be either class one or two customer) followed by a sequence of service times of accredited class-1 customers. During an accreditation interval, the time points at which customers become accredited occur according to a Poisson process with rate $\lambda_1(1 - b_2/b_1)$ (see [10], Lemma 4.2). Thus, the busy period of the queue can be divided into a sequence of accreditation intervals, which act as the effective service times from the prospective of class-2 customers.

Adopting the same idea as in the derivation of a busy period for M/G/1 queues, the LST of the distribution of the duration of an accreditation interval would satisfy the functional equation

$$\widetilde{\Gamma}(s) = \widetilde{B}(s + \lambda_1(1 - b_2/b_1)(1 - \widetilde{\Gamma}(s))), \tag{2.28}$$

which finally led to the derivation of the LST of waiting time distributions for the $M/G/c$ linear (non-preemptive) APQ.

2.4.1 Some fundamental results in APQ framework

In the following chapters we will extensively refer to the LST of the waiting time distributions in a 2-class priority APQ for c server cases. Hereby, we briefly present the necessary derivations required in future chapters in this regard.

It is also important to mention that in this thesis wherever we talk about APQs we mean non-preemptive APQs unless we state otherwise.

Lemma 1 *In a linear APQ under $M/M/c$ discipline, let $b_1 = 1$ and $b_2 = b$ and let $\Phi(s)$ be the LST of the service distribution. The LST of the stationary waiting time distribution for the class-2 customers is given by:*

$$\tilde{w}_2(s) = \tilde{w}(s + \lambda_1(1 - b)(1 - \tilde{\eta}_c(s))), \tag{2.29}$$

where $\tilde{\eta}_1(s)$ is the LST of the duration of the busy period for accredited customers and $\tilde{w}(s)$

represents the LST of the waiting time in an $M/M/c$ queue, which could be obtained by:

$$\begin{aligned}\tilde{\eta}_c(s) &= \Phi(s + \lambda_1(1-b)(1-\tilde{\eta}_c(s))) \\ &= c\mu/(c\mu + [s + \lambda_1(1-b)(1-\tilde{\eta}_c(s))]) \\ &= \frac{(s + \mu c + \lambda_1(1-b)) - \sqrt{(s+\mu c+\lambda_1(1-b))^2 - 4\lambda_1(1-b)\mu c}}{2\lambda_1(1-b)}\end{aligned} \tag{2.30}$$

and,

$$\tilde{w}(s) = [1 - C(A,c)] + C(A,c)[(c\mu - \lambda)/(c\mu - \lambda + s)] \tag{2.31}$$

where $C(A,c) = \frac{A^c}{c!(1-\rho)}/(\sum_{i=0}^{c-1}\frac{A^i}{i!} + \frac{A^c}{c!(1-\rho)})$ is the probability that an arriving customer finds all of the c servers busy in an $M/M/c$ queue with $A = \lambda/\mu$.

Proof For a detailed proof see [12].

Therefore, if $c = 1$ and thus $C(A,c) = 1 - \rho$, we have:

$$\begin{aligned}\tilde{w}_2(s) &= \tilde{w}(s + \lambda_1(1-b)(1-\tilde{\eta}_c(s))) \\ &= \frac{(1-\rho)(s + \lambda_1(1-b)(1-\tilde{\eta}_1(s)))}{s - (1-\tilde{\eta}_1(s))(\lambda_1 b + \lambda_2)}\end{aligned} \tag{2.32}$$

And, if $c = 2$ we will obtain:

$$\begin{aligned}\tilde{w}_2(s) &= (1 - C(A,2)) + C(A,2)\frac{2\mu(1-\rho)}{2\mu(1-\rho) + s + \lambda_1(1-b)(1-\tilde{\eta}_2(s))} \\ &= (1 - C(A,2)) + C(A,2)\tilde{w}_2^{(+)}(s)\end{aligned} \tag{2.33}$$

where $C(A,2) = \frac{2\rho^2}{1+\rho}$ and the superscript $(+)$ refers to the positive waiting times when the arriving customer finds the server(s) busy.

Theorem 1.1 *In a linear APQ with 2 servers, $c = 2$, if $b_1 = 1$ and $b_2 = b = 0$, the waiting time distribution for class-2 customers will be similar to a classical priority queue.*

Proof In Kella & Yechiali (1985) [60], the LST of the waiting time of a class-k customer in a priority $M/M/c$ queue has been derived. Letting $k = 2$ we obtain:

$$\begin{aligned}\tilde{w}_2^p(s) &= (1 - C(A,c)) + C(A,c)\frac{2\mu(1-\rho)(1-\tilde{\eta}_2(s))}{s - (1-\tilde{\eta}_2(s))\lambda_2} \\ &= (1 - C(A,c)) + C(A,c)\tilde{w}_2^{p(+)}(s)\end{aligned}$$

where the superscript p in our notation simply refers to a priority queue.

From (2.30), we have

$$\tilde{\eta}_2(s) = \frac{2\mu}{2\mu + s + \lambda_1(1-\tilde{\eta}_2(s))}$$

for $c = 2$. By substituting it in (2.33) and simplifying the expression, we will have $\tilde{w}_2^p(s)$ as it it defined and this completes the proof.

2.4.2 Affine Accumulating Priority Queue (Affine APQ)

In all of the previous (*classical APQ*) studies, the initial priority of patients at the time of entrance is zero while the higher the priority of a patient, the greater the rate at which that patient accumulates priority.

However, in this study, we are interested to look at a situation when patients arrive with an initial class-dependent priority and earn credits as they wait. We intend to focus on the two class case in which, without loss of generality, we assume that only the higher priority class has an initial positive credit a.

Assume a single-server queue with Poisson arrivals and general service time distributions. Customers of class $i; i = 1, 2$ arrive at the queue as a Poisson process with rate λ_i. Upon arrival, a customer of class i, C_i, starts accumulating priority at rate b_i, where $b_1 > b_2$. Therefore, accumulated priorities of class 1 and 2 customers at time t_2 would be $(t_2 - \tau_1)b_1 + a$ and $(t_2 - \tau_2)b_2$ according to the priority function

$$q_i(t) = a_i + b_i(t - \tau_i) \tag{2.34}$$

respectively, when we assume that the customer C_i has entered the system at time τ_i, and a is the initial credit for class one customers (i.e. $a_1 = a$ and $a_2 = 0$).

Figure (2.1) plots the accumulated priorities of customers against time as a sample path of such processes. In this figure, we assume that $b_1 = 1$ and $b_2 = 0.5$ for illustrative purposes.

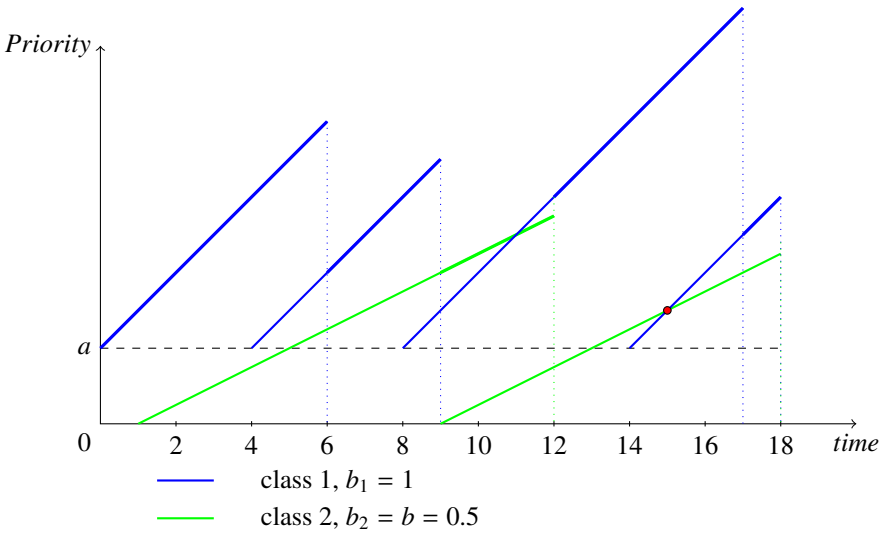

Figure 2.1: Accumulated priorities in a two class Affine APQ

In Figure (2.1) the arrival instants for class one customers are at points (0, 4, 8, 14) when the priority functions for this class are initiated, and departures occur at points (6, 9, 17, 18). Each heavy line is an indicator that the corresponding customer is in the service at that instant. In

this specific illustrated example, two lower class customers join the queue at times 1 and 9 when the latter is being overtaken by a class 1 customer at time 15.

Figure (2.2) however, demonstrates the Maximum Priority Process for the sample path of figure (2.1), superimposed on the priority functions.

Maximum Priority Processes in an Affine APQ are slightly different from the classical APQ, specifically as a result of the initial positive credit which class one customers gain upon arrival.

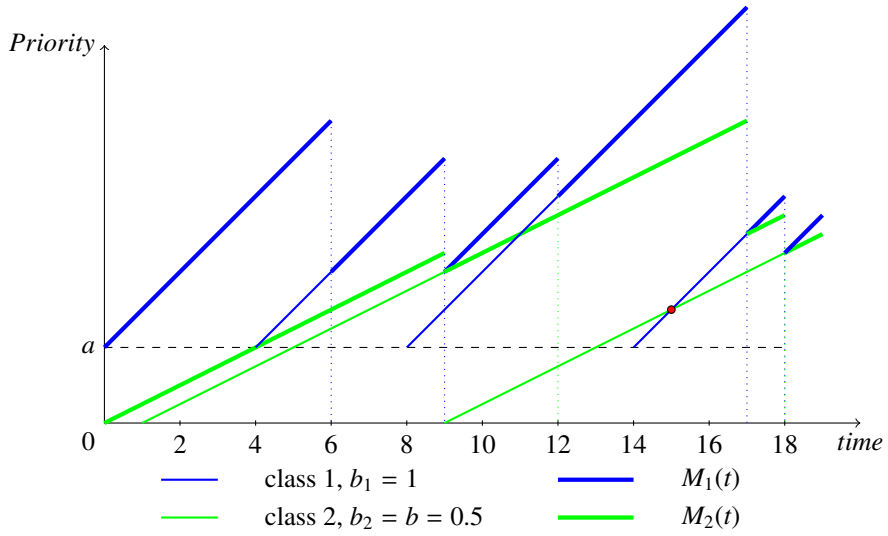

Figure 2.2: Maximum priorities in a two class Affine APQ

Definition 2.4.1 *Let $n(k)$ denote the arrival position of the kth customer to be serviced. The maximum priority process for the Affine APQ in the two-class case is a two-dimensional stochastic process $M(t) = (M_1(t), M_2(t)), t \geq 0$ which is defined as follows:*

1. *M(t)=(a,0) for all t corresponding to the idle periods.*
2. *For all t not corresponding to service commencements/completion instants, we have*

$$\frac{M_1(t)}{dt} = b_1 \quad \textit{and,} \quad \frac{M_2(t)}{dt} = b_2, \tag{2.35}$$

   where $0 \leq b_2 \leq b_1$.
3. *At the sequence of service completion times $\{\delta_k\}_{k=1}^{\infty}$,*

$$M_1(\delta_k) = \max\{a, q_\vee(\delta_k^-)\} \tag{2.36}$$
$$M_2(\delta_k) = \min\{M_1(\delta_k), M_2(\delta_k^-)\} \tag{2.37}$$

   where

$$q_\vee(\delta_k^-) = \max_{m\in\{n(k)+1,n(k)+2,\ldots\}} \{q_m(\delta_k^-)\} \tag{2.38}$$

   and $I\{A\}$ is the indicator function of the event A. Also, $q_m(\cdot)$ is the function defined in Equation (2.34).

Consequently, the 2-class affine APQ impacts the duration of the "accreditation cycles" which act as the APQs effective service times. The initial credit also changes the probabilistic nature of the delay cycles that a customer of a given class could encounter. Therefore, adjustments to the structure of accreditation intervals will be required, thus reformulating the busy period.

A special variant of APQ is called the "Delayed APQ". In this queue the lower priority class of customers spends some time, $d = a/b$, before they start to accumulate (positive) priority credit ($a_1 = 0$ and $a_2 = -a$). Such situations arise when it is felt that lower-priority customers waiting times are only of concern or interest once they reach a threshold, and hence only starts priority accrual from that point onward. In fact, all arrivals of high-priority would immediately overtake a class-2 customer until the latter had reached time d, subsequent to which the regular accreditation rate of $\lambda_1(1-b)$ as in the "classical APQ" would apply. In chapter 5 we show that at least for the two-class case, a full equivalence exists between the affine APQ and the "Delayed APQ".

As equivalent to the "Classical APQ", in Affine APQ, the accumulated priority credits of the customers still waiting are also distributed as Poisson processes according to the following theorem.

Theorem 2.4.1 *Let $t \in [0,\infty)$ and $\mathcal{M}(t) \equiv \sigma\{(M_1(u), M_2(u)), u \in [0,t]\}$ be the filtration generated by the maximum priority process up to time t in an Affine APQ, conditional on $\mathcal{M}(t)$:*

1) The accumulated priorities $\{q_k^i(t), k = 1,2,...\}$ of the customers still waiting from class i; $i = 1,2$ are distributed as independent Poisson processes with rate λ_1/b_1 on the interval $[a, M_1(t)]$ and with rate λ_2/b_2 on $[0, M_2(t)]$.

2) Let $M_2^{a+}(t) = \max\{M_2(t), a\}$ and $M_2^{a-}(t) = \min\{M_2(t), a\}$.
The accumulated priorities $\{q_k(t), k = 1,2,...\}$ of all customers still present in the queue are distributed as a Poisson process with rate zero on the intervals $[M_1(t), \infty)$, λ_1/b_1 on the interval $[M_2^{a+}(t), M_1(t))$, $I\{M_2(t) > a\} \cdot (\lambda_1/b_1 + \lambda_2/b_2)$ on the interval $[M_2^{a-}(t), M_2^{a+}(t))$ and λ_2/b_2 on the interval $[0, M_2^{a-}(t))$.

Proof 1) Let $a_1 = a$ and $a_2 = 0$ as the class-dependant initial credits. If there is no customer in service at time t, the statement of the theorem is trivially true. Otherwise, let $\tau < t$ be the time at which the current service commenced. The maximal priority of any class i customer queued at time τ was $M_t(\tau)$, which implies that a class i customer must have arrived at time $\tau - \frac{M_1(\tau)-a_i}{b_i}$. These customers who either were present in the queue with priority less than $M_i(\tau)$ or arrived in the queue in the interval (τ, t) have arrival instants which occurred according to Poisson processes with rate λ_i on interval $(\tau - \frac{M_i(\tau)-a_i}{b_i}, t]$. The priorities of these customers at time t can be calculated as $q_k^i(t) = a_i + b_i(t - \tau_i)$ which occur according to a Poisson process with parameter λ_i/b_i on interval $[a_i, M_i(\tau) + b_i(t-\tau)) = [a_i, M_i(t))$.

2) Proof is similar to Theorem 3.2 (2) in [10].

In other words, Theorem (2.4.1) states that there could be two possible scenarios regarding the positioning of $M_1(t)$, a and $M_2(t)$ in relation to each other. Figure 2.3 illustrates these two

situations. Figure 2.3 (A) presents one example case in which $M_2(t)$ is larger than a. In this scenario, as stated in the above theorem, the accumulated priorities $\{q_k(t), k = 1, 2, ...\}$ of all present customers in the queue is distributed according to a Poisson process with rate λ_2/b_2 on $[0, a)$, $\lambda_2/b_2 + \lambda_1/b_1$ on $[a, M_2(t))$, λ_1/b_1 on $[M_2(t), M_1(t))$.

In Figure 2.3 (B), we see an example where $M_2(t) < a$. In this scenario, the accumulated priorities $\{q_k(t), k = 1, 2, ...\}$ of all present customers in the queue is distributed according to a Poisson process with rate λ_2/b_2 on $[0, M_2(t))$, 0 on $[M_2(t), a)$, λ_1/b_1 on $[a, M_1(t))$.

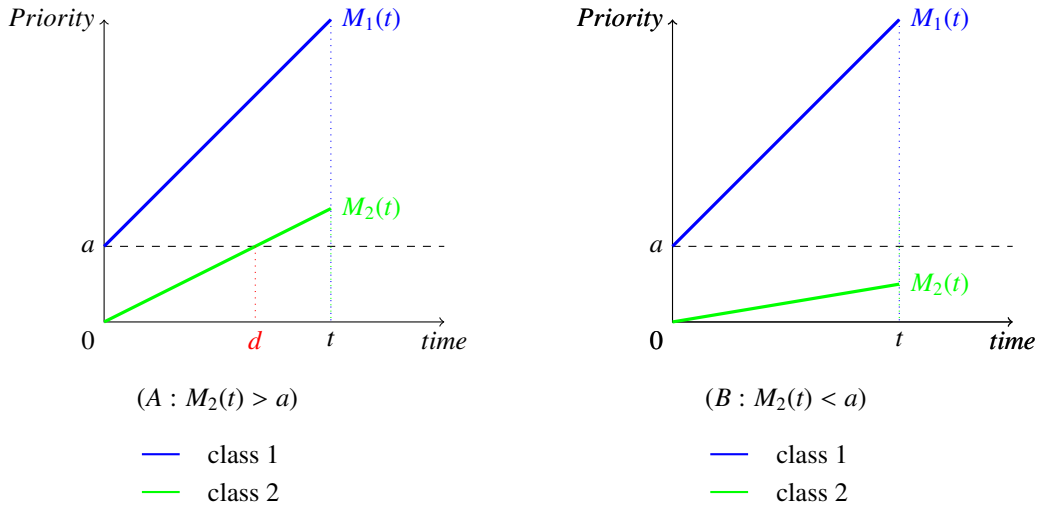

Figure 2.3: Accumulated priorities in an Affine APQ and Delayed APQ

2.5 Some elementary concepts

1. Many important and significant problems in engineering, the physical sciences, and the social sciences, when formulated in mathematical terms, require the determination of a function satisfying an equation containing derivatives of the unknown functions. Such equations are called **differential equations** [15].

There are different ways to classify differential equations. In ordinary differential equations only ordinary derivatives appear in the equations; whereas, in partial differential equations the partial derivatives will form the equation. The highest order of derivatives that appear in an equation will determines the order of the equation. The equation

$$F(x, u(x), u'(x), ..., u^{(n)}(x)) = 0 \tag{2.39}$$

is an ordinary equation of the nth order. Another main classification for differential equations are the class of linear and nonlinear equations. A differential equation is called linear if it is a linear function of the variables, otherwise it is called nonlinear. Therefore the general linear ordinary differential equation of order n is in the form of:

$$b_0(x)y^{(n)} + b_1(x)y^{(n-1)} + ... + b_n(x)y = g(x). \tag{2.40}$$

A solution for the ordinary linear equation of order 1 in the form of $y' + p(x)y = g(x)$ is [15]

$$y = \frac{1}{\mu(x)}[\int_a^x \mu(s)g(s)ds + C], \tag{2.41}$$

where

$$\mu(x) = \exp \int_a^x p(t)dt. \tag{2.42}$$

If we let $x = a$, then $y(a) = C$. Therefore, we can re-write the Equation (2.41) as $y = \frac{1}{\mu(x)}[\int_a^x \mu(s)g(s)ds + y(a)]$.

2. A collection of random variables, $X = \{X(t), t \in T\}$, is called a **stochastic process** when for each t in the index set T, $X(t)$ is a random variable. If T is a continuous set of values, the stochastic process is said to be a continuous-time process and if a countable or finite set, then we have a discrete-time process which would be referred to as X_n. Any realization of X is called a **sample path**.

The state space of a stochastic process is defined as the set of all possible values (or states) that the random variables $X(t)$ can take. If the state space is finite or countable, then we have a discrete-state process, often referred to as a chain. On the other hand, if the state space is a continuous interval (or a set of such intervals), we have a continuous state process [29].

The **Markov property** which is expressed analytically as

$$P\{X(t_{n+1}) = x_{n+1}|X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, ..., X(t_1) = x_1\} = P\{X(t_{n+1}) = x_{n+1}|X(t_n) = x_n\}, \tag{2.43}$$

states that the past history can be completely summarised in the current realisation. In other words, given the current state, future is independent of the past.

A stochastic process with Markov property is a **Markov Process**; and, a Markov process with with a discrete state space is often referred to as a **Markov chain**.

For more details about Markov chains see [29],[32] or Appendix A4 in [14].

Let $\mathcal{S}$ be the state space in a Markov process (on $\mathcal{S}$), for all $x \in \mathcal{S}$, the **transition kernel** $P(x, A)$ is defined as the probability of reaching the measurable set A from state x.

For all $x \in \mathcal{S}$, $p(x, y)$ is defined as a non-negative function such that $P(x, A) = \int_{y \in A} p(x, y)dy$. in this case for every x, $p(x, .)$ is a pdf, and

$$P(x, \mathcal{S}) = \int_{y \in \mathcal{S}} p(x, y)dy = 1.$$

Chapter 3

Discrete time Markov chain algorithm for short time predictions in an Emergency Department

Maryam Mojalal[1], Greg Zaric[2], David A. Stanford[1], and Alim Pardhan[3]

1: Department of Statistical & Actuarial Sciences, The University of Western Ontario, London, Canada

2: Richard Ivey Business School, The University of Western Ontario, London, Canada

3: Division of Emergency Medicine, Department of Medicine, McMaster University, Hamilton, Canada

3.1 Abstract

Short-run predictions of ED censuses are particularly important for efficient allocation and management of ED resources. In our study, we model ED changes and present estimations for short term (hourly) ED censuses at each time point. Considering the variation in arrival pattern and service requirements, we apply and compare three models which best describe our data. We provide hourly predictions up to 24 hours in a day and construct a numerical example to explain the effects of different possible interventions on preventing over-crowding in a system. We illustrate our approach using 22 months of data obtained from the ED of a large academic medical center in Ontario. Our three models will be validated and compared in accuracy and functionality based on MSE and correlation, R.

3.2 Introduction

Long wait times for health care services is an important health policy issue, as its consequences may well include adverse effects on patient health. Long wait times for just four procedures (joint replacement surgery, sight restoration, coronary artery bypass graft surgery and MRI scans) cost the Canadian economy an estimated $14.8 billion in 2007 according to the Report Card on Wait times in Canada [11]. Many factors contribute to long wait times in Emergency departments (EDs), such as a shortage of acute care bed capacity or limited community care resources.

As a result, clinically specified maximum wait-time benchmarks have been established to prioritize patients and implement appropriate treatment. For instance, the Canadian Triage and Acuity Scale (CTAS) [22] was introduced to provide benchmarks for Emergency care. Furthermore, several studies have tried to assist health care management by providing short or long term forecasts of the demand fluctuations on the system and its individual zones in the emergency department using statistical models.

Often overcrowding in an ED is described as a mismatch between patient demand for services and provider supply of resources [52]. Therefore, it is an obstacle to the timely delivery of health care to patients. It has also been linked to lower profitability, poor patient outcomes and higher operational costs [1]. One approach to alleviate problems associated with ED overcrowding is to forecast levels of demand for ED in advance in order to give health-care managers an opportunity to prepare for surges in demand and plan appropriate strategies.

The goal of this study is to accurately predict the hourly number of patients in the system, i.e., *census*, considering the dependency and seasonality aspects of census by using forecasting methods. In order to derive predictions for the near future, we introduce a Backward Algorithm in the proceeding section. Furthermore, we develop a regression model and Markov-chain based models to provide 24-hour forecasts for the near-future census at the ED of Hamilton University hospital. Finally, we validate our models on historical data using the "Sliding window" approach and compare their respective effectiveness based on Mean Square Error (MSE) and the coefficient of correlation (R). Time efficiency and interpretability of models are crucial factors considered in our study in order to make our model more applicable for ED administrators.

The remainder of this paper is organized as follows: the overview of the related literature and study setting is presented in the next section. The fourth section presents a brief description of our data set together with some preliminary data analysis and descriptive measures. In the fifth section we present our main developed algorithm in details. This algorithm serves as a building block for two of the forecasting models presented in section six. In section seven we explain how our developed model can potentially assist decision makers in an ED. The numerical data analysis and validation of the models by using real data are presented in the eighth section. Finally, the last section reviews the main contributions in this work and offers some final remarks including possible extensions to the study.

3.3 Overview of related literature

The formal presentation of an ED census model begins with arrivals. In 2013 Cote et al. [23] provided a tutorial and introduced ED medical directors to a range of straightforward regression-based forecasting models. Their work helped to predict the number of arrivals to an ED in support of strategic, tactical, and operational planning and activities. However, there are a variety of other models including various regression models in [23], time series [45, 57], simulations [52] or queueing analyses in the literature which model the arrival process.

According to Sharif [12], who provided a detailed literature on ED arrivals, a great deal of research has been performed towards *long-term* predictions of arrivals (i.e. daily, weekly, monthly or yearly) while less has been devoted to shorter time horizons. Early attempts at using time series models to predict arrivals in ED were made by Milner in 1988 [41], who used ARIMA models most widely. Among the various aspects considered in such studies are the addition of independent variables such as daily ambient temperature, humidity, air quality or holiday effects in [62], considering different acuity levels or exponential smoothing techniques. In their paper, Jones et al. [57] considered temporal relationships between ED demands and IP hospitals to develop a multi-variate time series models and apply them to their hourly data set. Finally Sharif in [12] used Generalized Linear Autoregressive Moving Average, GLARMA, models to forecast arrivals in ED using two years of data from an ED in southwestern Ontario. He has also provided a detailed review of a variety of linear models applied in the literature for different time horizons.

Considering the Input-throughput-output framework, modeling the discharge process will also be studied in this chapter. Available literature shows statistical tools have been applied in this regard as well [52]. However, there has been many studies which emphasize on the effect of workload, patient census and congestion on productivity and service rates [12].

In spite of this, less work has been devoted to census predictions and overcrowding warnings. The first attempt to develop an early warning system for ED overcrowding was done by Hoot et al. in 2006 [50] which used logistic regression and neural networks to predict ED ambulance diversion status one hour into the future. They later deployed a Discrete Event Simulation model to forecast ED crowding in [52]. Multivariate time series approach (vector Autoregressive) for hourly data [57], seasonal ARIMA, exponential smoothing and artificial neural network models for daily forecasts [56] [33], are the most frequent models used. Another study [24] demonstrated that workload and patient flow are linked via a simple input-output relationship and define two efficiency and congestion functions to model ED throughput.

Early work on Markov chain based prediction was done by Gabriel and Neumann in 1962 [49] in the area of statistical weather forecasting to predict daily rainfall for a single station. Fraedrich and Muller in 1983 [26] extended this model type to predict sunshine periods and probability of precipitation. In 1987 Fraedrich et al. [27] presented the theory for the linear combination of two independent predictive techniques useful for both short time weather prediction and long-term forecasting which lead to other studies such as [13] where Markov chain was combined with other predictive models. In these studies the transition probabilities of the Markov chain were mainly obtained empirically from relative frequencies in the histor-

ical data and final predictions were a weighted average of Markov chain predictions and other models.

A series of other studies are conducted in this area as well. Shamshad et al. [7] believe that due to their simplicity and because many natural processes are considered as Markov processes, Markov chains have become a popular tool for developing wind power prediction models based on time series analysis. In 2010 and 2015, [2] developed a wind power forecasting method based on discrete Markov chain models. In 2003, [18] proposed a short-term traffic flow forecasting method based on high order Markov chain theory.

In probability and queueing theory, Stanford et al. in 1983 [9] used the embedded stationary Markov chains to predict queue lengths in a $G/M/1$ queue. However in health care, Broyel et al. in 2010 [46] presented a Markov chain probability model that uses maximum likelihood regression to predict the expectations and discrete distributions of transient inpatient inventories. They expanded their work in 2011 [47] by employing a Markov decision process (MDP) to dynamically match hospital inpatient staffing to demand.

The analytical method which we develop here is based on Markov chains and will be applied to the short term census predictions in an emergency department. Our model accounts for the effect of calendar variables, seasonality of arrival and discharge processes and recent workload which has not been considered in previous studies. It also dynamically updates the predictions.

ED management can benefit from the resulting information by planning ED capacity, estimating required resources or efficient staffing schedules. One study has described management approaches that could be facilitated by statistical or other operational management techniques to reduce crowding [48]. Techniques such as the 1-bed-ahead strategy or more flexible staffing have been described as an effort to reduce over-crowding.

3.4 Data collection and study setting

For this study, two years of hourly data of 88000 adult ED visits were collected from a hospital in South Western Ontario. Patients who were registered and triaged between January 1, 2012 to December 31, 2013, were included in the study. Ethics approval for this study was obtained from the hospital. It is worth mentioning that the data set we worked on was a processed data received from a previous study as in [12].

From our data set hourly arrival and service counts was obtained. Descriptive analysis and model fitting were carried out using regression, time series and probabilistic methods. These models were compared to each other in terms of their ability to provide out-of-sample forecasts of ED census. All information about triage level, time of arrival, time of being seen by a physician and time of being discharged is available in our data set. Our data set consisted of individual time stamps for every single patient. Naturally, as often happens with clinical data collection, some missing data or badly reported data points were available. To achieve the complete information on missing value imputation methods and cleaning techniques which

had been applied to replace the missing data refer to [12].

Meanwhile, individual time stamps were aggregated to obtain arrival and discharge counts for each time interval. Arrivals in an emergency department are either in the form of walk-ins or by some kind of emergency transportation. The latter would be considered critical and by CTAS indicators classified as Resuscitation or Emergent, whereas walk-ins will be assigned an acuity level after being assessed by a triage-nurse. The most acute patients are treated immediately, while others need to wait until being called. Thus, patient received time has been considered as arrival time to the ED, and they are considered discharged from an ED when they are released home or admitted to other facilities of the hospital.

We had no direct access to actual ED census values in this study, though they play a crucial role for the modeling purposes. Therefore, we worked with the initial minimum queue length on January 1st 2012, so that no negative censuses are predicted. Then, by addition and subtraction of arrivals and discharged patients we calculated the census for the rest of data points.

Hourly data will be aggregated later to form 3-hour time blocks for the benefit of algorithm runtime. This strategy also helps to avoid zero counts which are dominant in the reported hourly ED arrival data. As a result, there will be 8 blocks during each 24 hour period starting from midnight on each day. More information about the clinical setting and available resources is available in [12]. Furthermore, according to the same study, the number of patients who had left the ED without being seen by a physician was very small; therefore the information of those patients was removed from the data set.

3.4.1 Preliminary analysis and descriptive graphs

The Patient flow through an ED can be divided into the following three parts: input, throughput and output. For a given discrete time point, n, the ED census can be represented by:

$$Q_{n+1} = Q_n + A_n - S_n, \qquad n = 1, 2, ..., 24, \tag{3.1}$$

where a discrete random variable, Q_n, is the census at time n, A_n and S_n are numbers of arrivals and discharges during the n^{th} interval respectively.

We define the nth interval to be an interval between the two time points n and $n+1$ (see Figure (3.1)). Also, for more convenience, we define $U_n = A_n - S_n$ as the *census increment at time n* in this study.

Figure (3.2) provides a first view of our data regarding the daily volume of arrivals in year 2012. It can be concluded from the graph that there is no readily evident trend available around the mean arrivals line over such a long period of time.

It would be however interesting if we examine the overall trends of arrival and discharge volumes. Figure (3.3) displays average volume by hour over the weekly cycle.

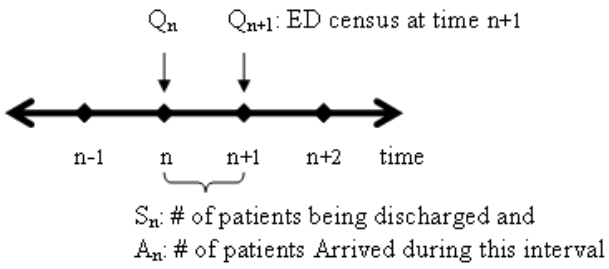

Figure 3.1: Variable description.

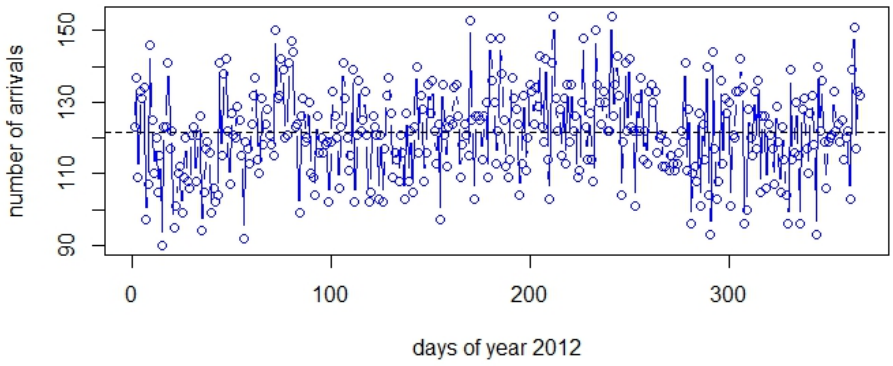

Figure 3.2: Daily arrival volume in year 2012.

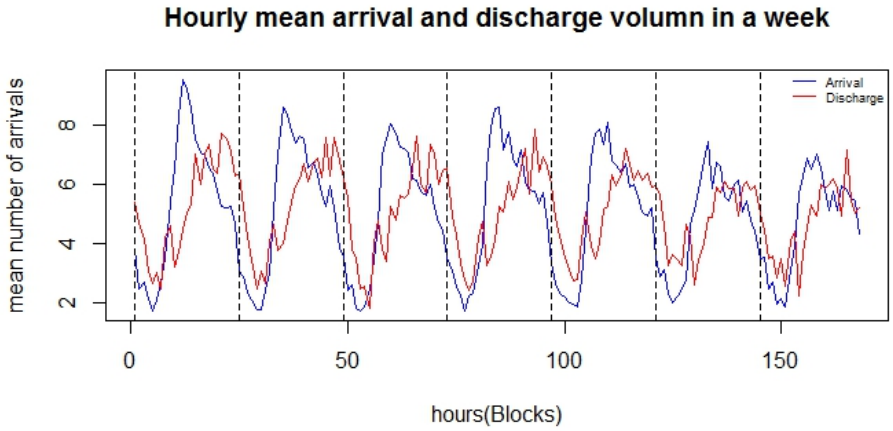

Figure 3.3: Mean number of arrival and discharge in 168 hours of a week.

This graph clearly suggests a daily seasonality over 24 hours of a day. The Highest arrival levels happen between 11 am and 4 pm. Then the volume of arrivals gradually decreases and reaches the minimum level between 3 and 7 in the mornings. This fact can be seen in more detail in figure (3.4). Furthermore, one can observe that Mondays have the highest arrival and service volume while Saturdays and especially Sundays have the lowest ones.

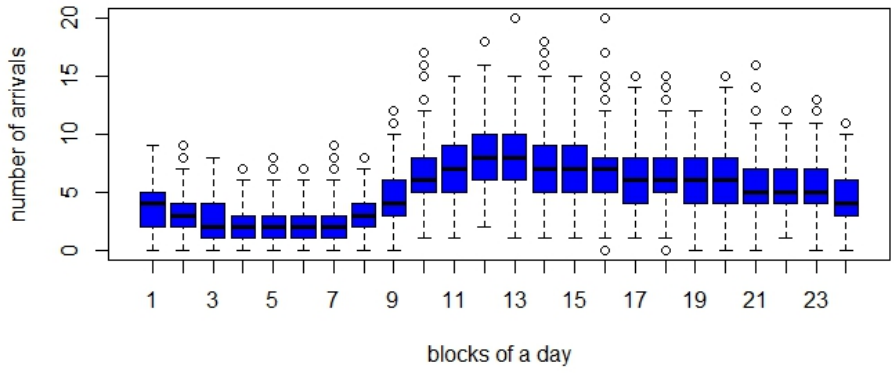

Figure 3.4: Number of hourly arrivals in 2012.

Another informative graph is the monthly graph as in figure (3.5). Fluctuations of both processes have been illustrated in this graph, where March, July and August have the highest number of arrivals and discharge volume in our ED.

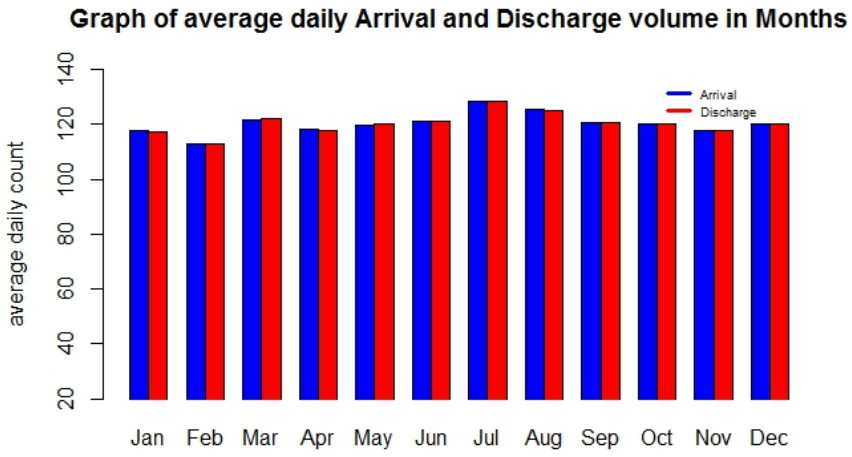

Figure 3.5: Number of arrival and discharge in months of 2012.

Fluctuations of the total census versus blocks of the day is another important graph in this study. In figure (3.6) we can see how the number of patients in ED gradually decreases until

8 am, and as it gets closer to noon there are more patients waiting; until finally from around 6 pm it declines with a smooth slope.

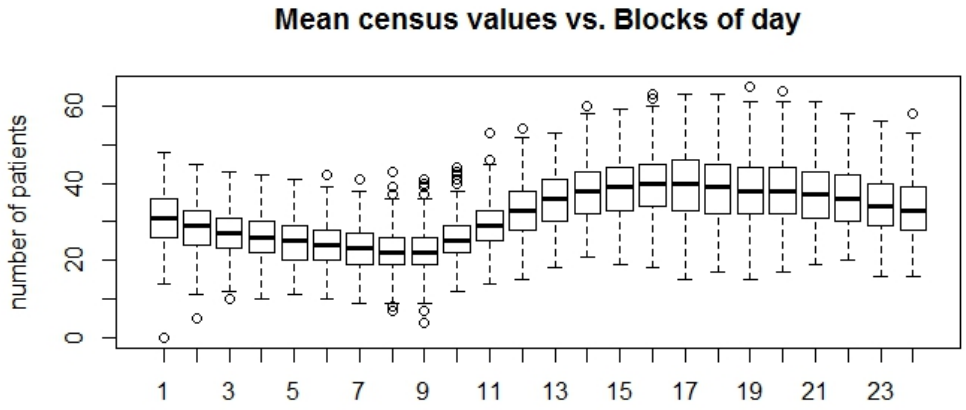

Figure 3.6: Box plot of number of patients in days of 2012.

Figure (3.7) illustrates the empirical distribution of the census increments in the n^{th} interval, U_n, based on the available data for four selected time points. These graphs suggest that the distributions have different means and variances.

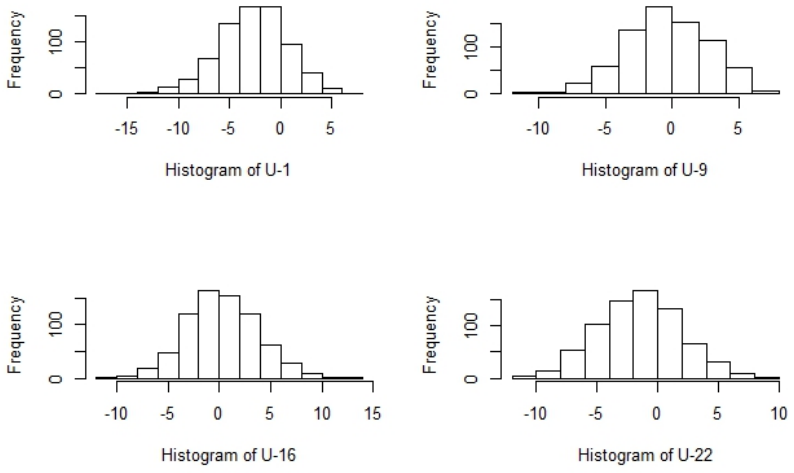

Figure 3.7: Histogram of U's at times 1, 9, 16 and 22.

Having gained an appreciation for the empirical nature of arrivals and discharges of the available data set here, in the next section we present models for making short term forecasts, which will be gained by examining the general trend of these processes.

3.5 Backward Discrete Time Markov Chain (DTMC) Algorithm

In this section we develop a Discrete Time Markov Chain (DTMC) probability formulation in the form of a recursive (backward) algorithm that captures the short term fluctuations of an ED. For that, let's recall from the previous section the general relation which describes dynamics of the census level where we had

$$Q_{n+1} = Q_n + A_n - S_n = Q_n + U_n, \tag{3.2}$$

where $U_n = A_n - S_n$ and $andS_n \leq Q_n + A_n$. This equation can even be applied recursively to state that the current census is simply the initial census level plus the difference between all arrival and discharges during the elapsed time.

$$Q_{n+1} = Q_k + \sum_{i=k}^{n} U_i. \tag{3.3}$$

Let $C_n(u)$ be the probability that the n^{th} increment, U_n, equals u. Thus, the distribution of U_n (under the assumption that Arrival and service processes are independent of the census level) for $u \geq 0$, is derived as

$$\begin{aligned} C_n(u) &= p[U_n = A_n - S_n = u] \\ &= \sum_{s=0}^{\infty} p(A_n - S_n = u | S_n = s) p(S_n = s) \\ &= \sum_{s=0}^{\infty} p(A_n = u + s) p(S_n = s). \end{aligned}$$

Since A_n and S_n, unlike U_n, are non-negative random variables, we need to re-write the foregoing derivations for $C_n(u)$ as

$$C_n(u) = \begin{cases} \sum_{s=0}^{\infty} p(A_n = u + s)\ p(S_n = s) & \text{if } u \geq 0, \\ \sum_{a=0}^{\infty} p(A_n = a)\ p(S_n = a - u) & \text{if } u < 0. \end{cases} \tag{3.4}$$

Thus, the probability distribution of the n^{th} increment can be viewed as the *convolution* of the number of arrivals and the negative of the number of discharges.

Let $p_{ij}^{(k,n)} = p(Q_n = j \mid Q_k = i)$ be the probability of transition from state i in the census at time k to state j at time n in a non-homogeneous process. Viewing the census process as a Markov chain, by means of the Chapman-Kolmogorov equation, this probability can be expanded as

$$p_{ij}^{(k,n)} = \sum_{l=0}^{\infty} p_{il}^{(k,k+1)} p_{lj}^{(k+1,n)}. \tag{3.5}$$

From (3.5) we get

$$\begin{aligned} E\{Q_n|Q_k = i\} &= \sum_{j=0}^{\infty} j p_{ij}^{(k,n)} \\ &= \sum_{j=0}^{\infty} j \sum_{l=0}^{\infty} p_{il}^{(k,k+1)} p_{lj}^{(k+1,n)} \\ &= \sum_{l=0}^{\infty} p_{il}^{(k,k+1)} \sum_{j=0}^{\infty} j p_{lj}^{(k+1,n)} \\ &= \sum_{l=0}^{\infty} p(U_k = l - i) E\{Q_n|Q_{k+1} = l\}. \end{aligned} \tag{3.6}$$

This algorithm works backward in time, starting from $k = n - 1$ where $E\{Q_n|Q_n = i\} = i$, recursively obtaining the predicted censuses for earlier time points.

In fact, the conditional census distributions for lag $(n - k)$ can be written as mixtures of the conditional census distributions for lag $(n - k - 1)$, so that we immediately find

$$P\{Q_n \leq h|Q_k = i\} = \sum_{l=0}^{\infty} p(U_k = l - i) P\{Q_n \leq h|Q_{k+1} = l\} \tag{3.7}$$

and,

$$E\{Q_n^2|Q_k = i\} = \sum_{l=0}^{\infty} p(U_k = l - i) E\{Q_n^2|Q_{k+1} = l\}. \tag{3.8}$$

Therefore, we readily find the variance which is,

$$Var\{Q_n|Q_k = i\} = E\{Q_n^2|Q_k = i\} - E^2\{Q_n|Q_k = i\}. \tag{3.9}$$

Since this algorithm enables us to obtain the probability distribution of the census, Q_n, along with the point estimations, we are able to use the percentiles of this distribution to find the required confidence intervals.

We illustrate the operation of this recursive scheme below, for the case of predictions for the midnight census.

I. Example: One to three-step prediction for midnight census

Letting $p_{ij}^{(23,24)} = p(Q_{24} = j \mid Q_{23} = i)$ denote the probability for the number of patients in the system at midnight (12:00 am) given i patients in the ED at 11:00 p.m., we immediately obtain

$$E\{Q_{24}|Q_{23} = i\} = \sum_{j=0}^{\infty} j p_{ij}^{(23,24)} = \sum_{l=0}^{\infty} p_{il}^{(23,24)} E\{Q_{24}|Q_{24} = l\} = \sum_{l=0}^{\infty} p(U_{23} = l - i) l. \quad (3.10)$$

Similarly, a prediction for midnight given i patients present at the ED at 10:00 is

$$E\{Q_{24}|Q_{22} = i\} = \sum_{r=0}^{\infty} p_{ir}^{(22,23)} E\{Q_{24}|Q_{23} = r\} = \sum_{r=0}^{\infty} p(U_{22} = r - i) E\{Q_{24}|Q_{23} = r\}. \quad (3.11)$$

This specifies a two-step prediction. However, a one-step prediction $E\{Q_{23}|Q_{22} = i\}$ can also be made similar to our first example. Likewise,

$$E\{Q_{24}|Q_{21} = i\} = \sum_{k=0}^{\infty} p_{ik}^{(21,22)} E\{Q_{24}|Q_{22} = k\} = \sum_{k=0}^{\infty} p(U_{21} = k - i) E\{Q_{24}|Q_{22} = k\}. \quad (3.12)$$

To produce estimates for the ED census using (3.4)-(3.6), we consider three methods for estimating the U_ns:

(1) A purely probabilistic model using Markov chains; in this approach which will be referred to as "the empirical approach", the empirical distribution of each U_n will be substituted in Equation (3.6),

(2) Directly modeling and forecasting the U_ns with time series models and applying (3.2) to obtain census predictions based upon a purely statistical approach, and

(3) Modeling the arrival and service processes via parametric regression models. Then deriving the U_n distributions,$C_n(u)$, using Equation (3.4) which we refer to as a hybrid model.

We discuss each of these three methods below in turn.

3.6 Models for making forecasts

The importance of short time predictions for an Emergency department was discussed in the previous section. In this chapter, we introduce three models that fit our data best. We explain some theoretical background wherever required, study their related assumptions and provide some numerical applications. Two of these models will use the backward algorithm developed in the previous section, and one other is a completely statistical model.

3.6.1 The Empirical Approach

We assume that the n^{th} increment's ,U_n, distribution can be approximated empirically from the historical data, and that the increments are independent of the current workload and calender variables. In this approach the empirical distribution is written as:

$$P(U_n = i) = \frac{\text{count number of } U_n \text{ in data set equal to } i}{\text{count number of all } U_n} \qquad \text{for all } n \in \{1, \ldots, 24\}$$

This provides us with an estimate for the probability values at each time point to substitute for the $p(U_k = l - i)$ in Equation (3.6). As a result, we are able to predict the census any step forward from a given time point.

Confidence Interval

Since our algorithm enable us to have a prediction for the whole distribution of the census in the near future, in order to identify the lower and upper bounds within which 95 percent of our data lies, we just need to identify the 2.5 and 97.5 percentiles of that distribution. Hence, the lower and upper bounds would be $L = \inf\{h : P(Q_n \leq h \mid Q_k = i) \geq 0.025\}$ and $U = \sup\{h : P(Q_n \leq h \mid Q_k = i) \leq 0.975\}$ respectively.

Note that we employ Equation (3.7) to obtain the c.d.f and required percentiles for each distribution.

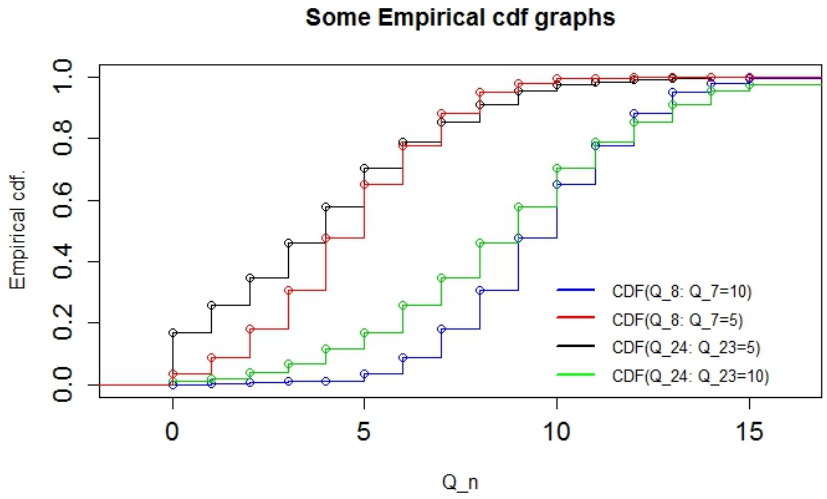

Figure 3.8: Empirical c.d.f function.

Figure (3.8) compares four empirical $c.d.f$'s one hour transition given 10 or 5 patients present at 7:00 am or 11:00 pm. The blue and green lines demonstrate the difference between morning hours with higher arrivals and night hours with higher service rates respectively. On the other hand, the difference between blue and red lines shows the effect of different initial queue length.

Once the c.d.f has been obtained, we are able to find the confidence interval for estimated values as shown in Figure (3.9) as an example.

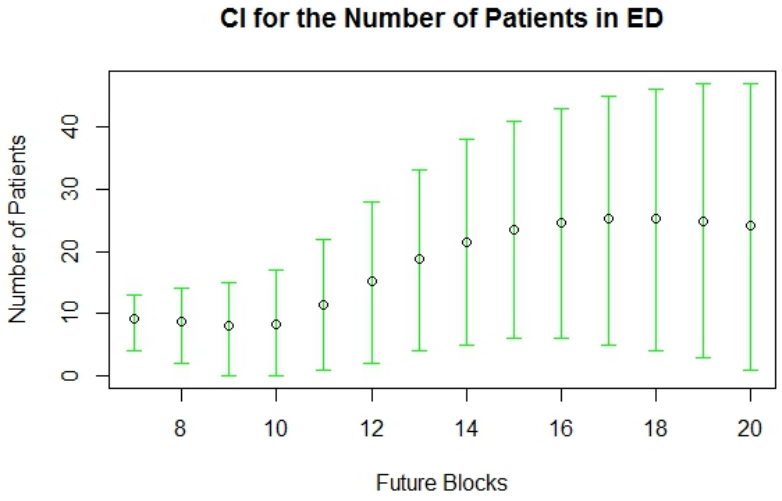

Figure 3.9: Confidence Intervals for census predictions.

In Figure (3.9), we have assumed that 10 patients were at an ED at time 6:00 am. The hourly predictions are displayed as round circles and the green lines represent the corresponding 95% confidence intervals.

3.6.2 Numerical implementations of the first model

Since our algorithm for census prediction works backward at each step, initially a recursive function was used in R which calls $E\{Q_n|Q_{k+1} = l\}$ to get $E\{Q_n|Q_k = i\}$ and so on. However, the recursive function turned out to be very time consuming, so an alternate algorithm was implemented. This function starts from $E\{Q_n|Q_n = i\}$; for the next step it finds $E\{Q_n|Q_{n-1} = i\}; i = 0, 1, ..., \max$ where $(n - 1 > k)$ and saves it in a vector whose length is, $max = i + \sum_{j=k}^{n-1}(\max u_j)^+$. Whenever each of these values is required in the next steps, the algorithm uses the vector it has saved from the previous step, which reduces the needed computation time a great deal.

Table 3.1: An example to demonstrate steps of the algorithm in R function

0	1	2	...	max
$E\{Q_7 \mid Q_7 = 0\}$	$E\{Q_7 \mid Q_7 = 1\}$	$E\{Q_7 \mid Q_7 = 2\}$	...	$E\{Q_7 \mid Q_7 = max\}$
$E\{Q_7 \mid Q_6 = 0\}$	$E\{Q_7 \mid Q_6 = 1\}$	$E\{Q_7 \mid Q_6 = 2\}$	...	$E\{Q_7 \mid Q_6 = max\}$
$E\{Q_7 \mid Q_5 = 0\}$	$E\{Q_7 \mid Q_5 = 1\}$	$E\{Q_7 \mid Q_5 = 2\}$	...	$E\{Q_7 \mid Q_5 = max\}$
$E\{Q_7 \mid Q_4 = 0\}$	$E\{Q_7 \mid Q_4 = 1\}$	$E\{Q_7 \mid Q_4 = 2\}$	...	$E\{Q_7 \mid Q_4 = max\}$

Table (3.1) presents the steps of our algorithm which finds $E\{Q_7 \mid Q_4 = 1\}$. More details can be found in the algorithm description in table (3.2).

Table 3.2: Backward Algorithm for census predictions

Algorithm description
1 : Derive the empirical U distribution as a function from data set
2 : Define a function to derive the maximum value of U_n values for each n where $n \in \{1, \ldots 24\}$
2 : and call it $maxU()$
3 : Define a function to take last.time, first.time, count as inputs
4 : **if** $last.time - first.time = 0$ **then**
5 : **return** $count \leftarrow count$
6 : **else**
7 : **if** $(last.time - first.time > 0)$ **then**
8 : $loopm \leftarrow last.time - first.time$
9 : **else**
10 : $loopm \leftarrow (last.time - first.time) + 8$
11 : $max.engaged.us \leftarrow max(sapply(1:8, maxU)) * loopm$
12 : $rmax \leftarrow max.engaged.us + count$
13 : $m \leftarrow vector(0, ..., rmax)$
14 : **for** $t = 1$ to $loopm$ **do**
15 : **if** $last.time - t > 0$ **then**
16 : $timet \leftarrow last.time - t$
17 : **else**
18 : $timet \leftarrow 8 + last.time - t$
19 : **for** $k = 0$ to $rmax$ **do**
20 : $S \leftarrow 0$
21 : **for** $r = 1$ to $rmax$ **do**
22 : $S \leftarrow P(U_{r-n} = timet) * m[r] + S$
23 : $S \leftarrow P(U_{0-n} = timet) * m[1] + S$
24 : $new - m[k] \leftarrow S$
25 : $m \leftarrow new - m$
26 : **return** $count \leftarrow m[count + 1]$

The necessary steps of an algorithm to find the census prediction appear in Table (3.2). Lines 7 to 10 find the number of steps ahead we seek to predict. In case the future time point is in the next day, we need to add 8 steps to the subtraction of future time from current time to avoid negative lags. Line 12 finds the upper limit of summation to avoid infinite loops in Equation (3.6), while later loops are to calculate all the one, two up to required step predictions into the future and to save them in a vector.

Table (3.3) displays numerical examples for some typical initial and predicted census values (4 and 8 hours into the future) where k is the initial time epoch, i represents initial census level and

n is the time we are interested to make predictions for. For example, the algorithm's predicted value for 4:00 pm, given 10 patients present at 8:00 am is 26.

Table 3.3: An Example of prediction for different i, k and n

	$\mathbf{E\{Q_n \mid Q_k = i\}}$					
k	i = 0		i=10		i=40	
1	n=4 1	n=8 2	n=4 5	n=8 4	n=4 34	n=8 31
4	n=8 2	n=12 10	n=8 7	n=12 14	n=8 37	n=12 43
8	n=12 9	n=16 18	n=12 16	n=16 26	n=12 46	n=16 55
16	n=20 3	n=24 4	n=20 10	n=24 8	n=20 39	n=24 36

3.6.3 Regression with autoregressive moving averages (ARMA) errors

The second approach is to model the fluctuations of U_n's over time. We developed a model based on ordinary regression and time series Autoregressive Moving Average (ARMA) models.

There is extensive research that supports the application of time series analysis to improve forecasting in health care. ARIMA models have been used to predict ED patient volume across short- and long-term time horizons [1520]. Several studies such as Marcilio et al. (2013), have compared different regression models to ARIMA models and have made conclusions about which model is a better fit based on their available data.

We fitted an ordinary regression to the increments, U_n to find significant calender variables, and tested the residuals for any possible trend. The methodological framework used in this study is presented as follows:

- Plot U_n data over time as a time series,
- Fit an ordinary regression model and save the residuals,
- Fit several models to the obtained residuals and estimate model parameters using dependency measured, e.g., Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)
- Identify best models using fit criteria, e.g., Akaikes Information Criterion (AIC), Bias Corrected AIC (AICc), and Bayesian Information Criterion (BIC)
- Apply diagnostic tools to determine how well the models fit census data, e.g., Plot of standardized residuals and their normal Q-Q plot, and
- Forecast n days during specified periods using an independent dataset. Evaluate the accuracy of the forecasts provided by the models by using error statistics, e.g., Mean Absolute Percentage

It is possible (as in our case) that the errors (residuals) of a regression model have a time series structure or have high autocorrelation which violates the initial assumptions of the ordinary linear regression model, where residuals have to be independent. Thus, in order to avoid biased and inflated coefficient estimations, we can adjust the regression coefficient estimations and standard errors.

In this procedure the underlying model is called a regression with auto correlated errors [59]. That is, consider the model to be given by:

$$\mathbf{y} = X\beta + \mathbf{e}, \tag{3.13}$$

where $\mathbf{y}$ is an $n \times 1$ vector, X is the $n \times r$ Regression covariate matrix (fixed input), β is an $r \times 1$ vector of regression parameters and n is the number of observations. The $n \times 1$ error vector, $\mathbf{e}$, is a process with come covariance implying that the error terms are correlated (i.e. are not white noise, w_t).

Now, if we were working with a pure $AR(p)$ error we could write:

$$e_t = \Phi^{-1}(B)w_t, \tag{3.14}$$

where $\Phi(B)$ is the linear transformation that, when applied to the error process, produces the white noise w_t. Notice that the model has two error terms here; the error from the regression model which we denote by $\mathbf{e}$ and the error from the ARIMA model which we denote by $\mathbf{w}$. Only the ARIMA model errors are assumed to be white noise.

Initial investigations

Figure (3.10) provides a first view of increments. It shows values of U_n versus time of the day for 40 days, and suggests U_n is a stationary process. However, the pattern across lags in Figure (3.11) are multiples of 8 (as we have 8 blocks in a day), which indicates daily seasonality. This figure illustrates the Auto correlation function (ACF) and Partial auto correlation functions for U_n.

We divided the data set to two separate sections: Training data and Test data, where the latter consists of 20 percent of whole data set.

Fitting a linear regression model to the training data set revealed no significant monthly effect and the only important variables were blocks of day which are defined as $I_{block_n}, n = 1, 2, ..., 8$ and Monday effect (I_{Monday}). Figure (3.12) illustrates the autocorrelation of the residuals of the regression model.

Therefore, in the next step the ARMA $(7, 0, 2)$ was fitted to the residuals of linear regression model. The residual diagnostic of the fitted ARMA model could be viewed in in figure (3.13) which shows no significant deviation from assumptions.

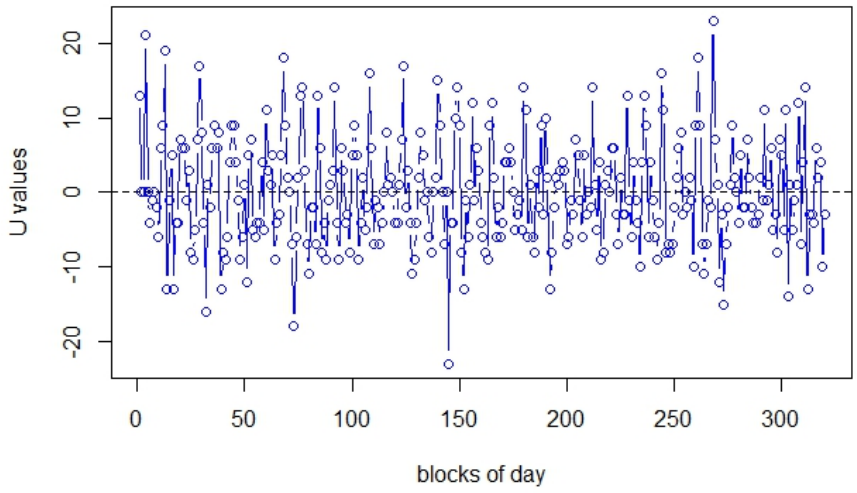

Figure 3.10: The hourly values of U_n in 40 days.

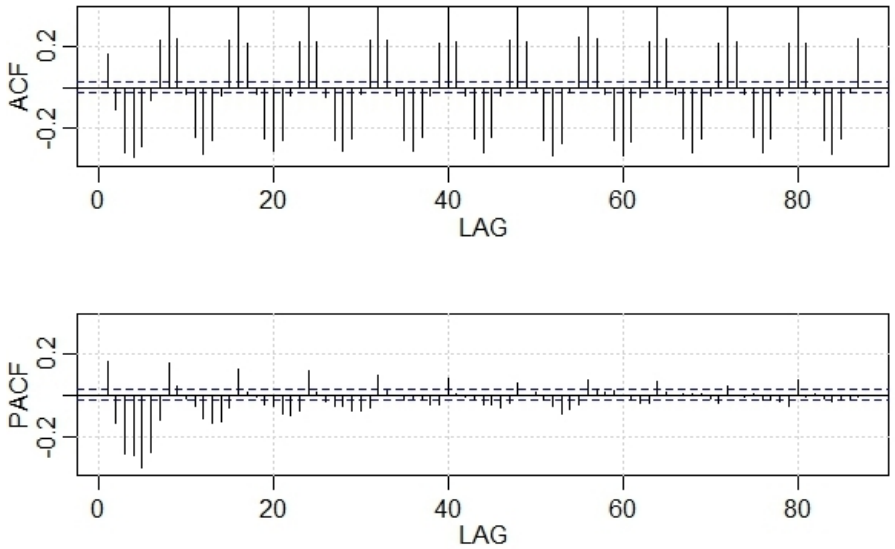

Figure 3.11: The ACF and PACF of U_n

Finally, the resulting model which could be written as

$$\begin{aligned} U_t = &-6.19 + 3.04 \times I_{block2} + 4.78 \times I_{block3} + 16.46 \times I_{block4} + 11.80 \times I_{block5} + 6.22 \times I_{block6} + \\ &3.92 \times I_{block7} + 2.38 \times I_{block8} + 0.85 \times I_{Monday} + 1.35 \times e_{t-1} - 0.51 \times e_{t-2} + 0.02 \times e_{t-3} + \\ &0.05 \times e_{t-4} - 0.06 \times e_{t-5} + 0.03 \times e_{t-6} + 0.02 \times e_{t-7} + w_t - 1.62 \times w_{t-1} + 0.62 \times w_{t-2} \end{aligned} \tag{3.15}$$

is being validated on our test data set. Figure (3.14) demonstrates the forecast values and confidence intervals related to each forecast value.

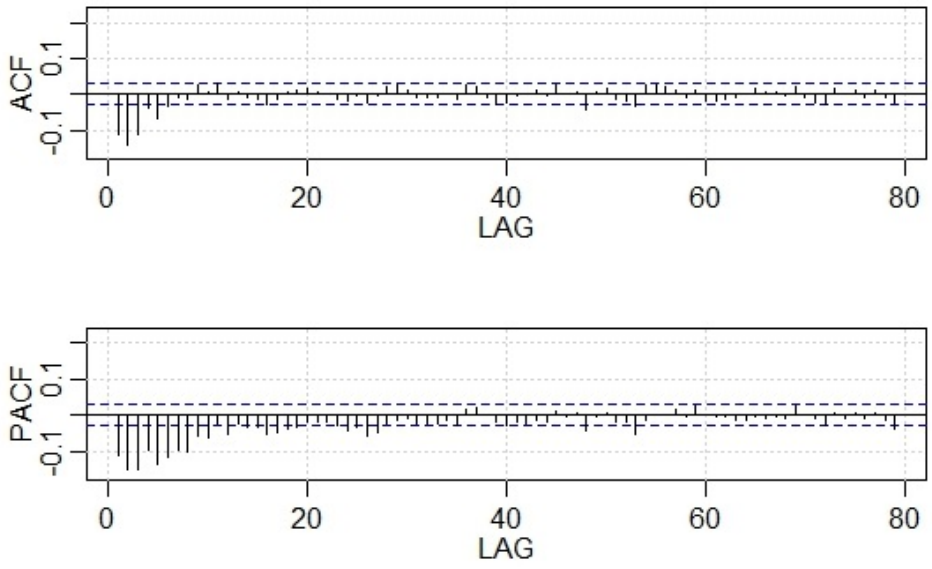

Figure 3.12: ACF and PACF of the residuals of the linear regression model.

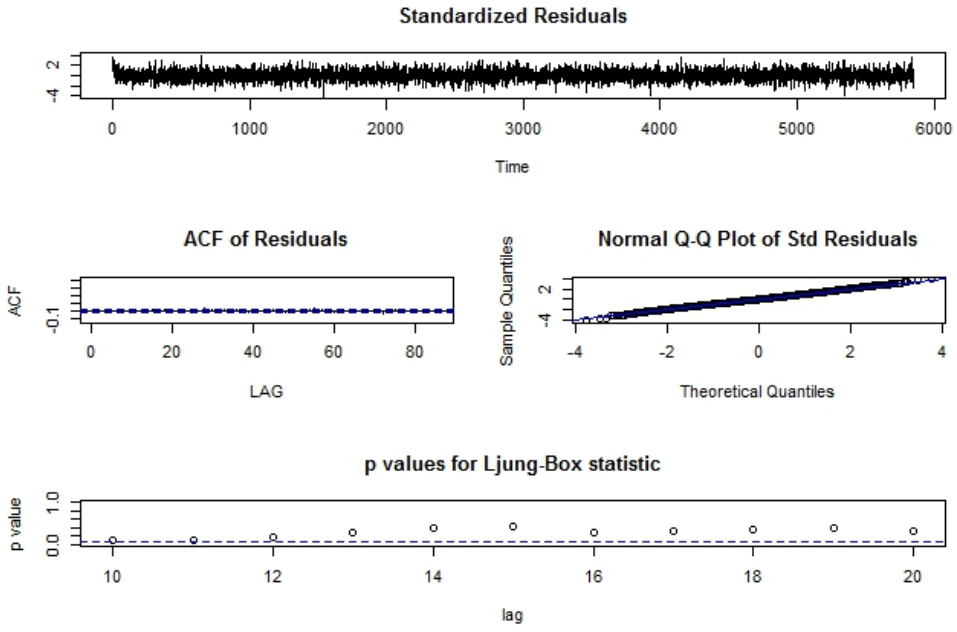

Figure 3.13: Residual diagnostic for the fitted ARMA model.

Having forecasts for U_n values, we can now derive forecasts for ED census directly using equation (3.2).

Figure (3.15) shows one, three, five and seven step forecasts of ED census (in blue) versus the real census volumes (in green). Real values have been available from our test data while forecasted values are derived from the regression with ARMA errors model.

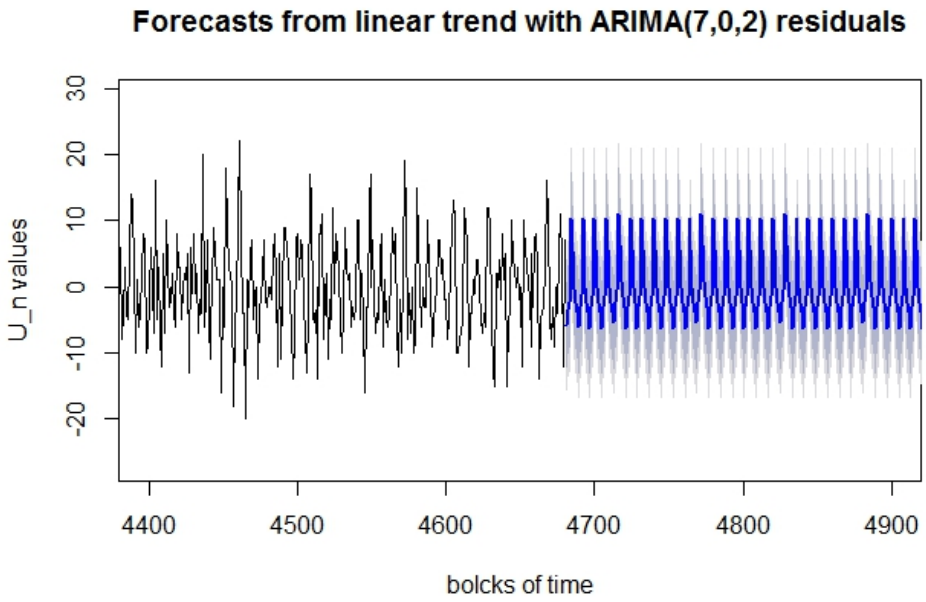

Figure 3.14: Residual diagnostic analysis of U_n time series.

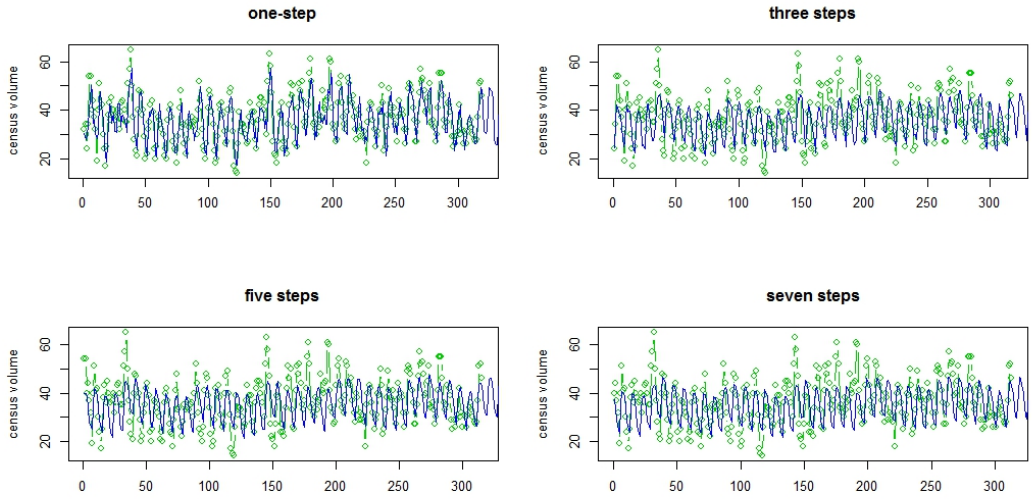

Figure 3.15: One to seven steps forecasts of ED census.

3.6.4 Hybrid Model: Parametric discrete Markov chain approach

This section presents a DTMC probability formulation that captures the short term effects of census fluctuations in an Emergency Department. Because service (discharge) and arrival rates are considered unknown by hour of the day, the probability model will be combined with a statistical approach to predict arrival and discharge rates based on historical data. Output results to measure the adequacy of models have been presented subsequently.

Formulation of the model

Our third approach represents a hybrid of the previous two. As with the first method, we worked with a Markov-chain based recursion exploiting the relevant increment distributions. Here, however, the increment distributions are obtained as the difference between parametric arrival and discharge distributions. Each of the two parametric distributions have parameters that are regressed upon quantities from the preceding time interval.

For that purpose, we need to obtain the arrival and discharge probabilities at each time point. Specifically, we first fit a Poisson regression model to our arrival and discharge data. A previous study [12] on the same data set has studied and proved the effects of workload and calendar variables on both Arrival and Discharge processes. In the light of that, we derive a prediction for the distribution of increments and use a modification of equation (3.6) to derive the required forecasts for future census.

Since our arrival and service volume depend on workload in the system, we need to modify our recursion equation, Equation (3.6), to account for that too.

To begin with, we assume our census is a Markov process which depends only on one previous state:

$$P(Q_n = j \mid Q_k = i) = \sum_{l=0}^{\infty} P(Q_n = j \mid Q_{k+1} = l) P(Q_{k+1} = l \mid Q_k = i). \tag{3.16}$$

Furthermore, assuming the arrival and departure numbers follow Poisson distributions, λ_k and μ_k are the arrival and discharge rate for each time interval respectively. Therefore, the equivalent recursion prediction formula in this setting is written as:

$$\begin{aligned}
&E\{Q_n | Q_k = i, \lambda_k, \mu_k\} \\
&= \sum_{j=0}^{\infty} j \sum_{l=0}^{\infty} P(Q_{k+1} = l \mid Q_k = i, \lambda_k, \mu_k) P(Q_n = j \mid Q_{k+1} = l, \lambda_{k+1}, \mu_{k+1}) \\
&= \sum_{l=0}^{\infty} P(Q_{k+1} = l \mid Q_k = i, \lambda_k, \mu_k) \sum_{j=0}^{\infty} j P(Q_n = j \mid Q_{k+1} = l, \lambda_{k+1}, \mu_{k+1}) \\
&= \sum_{l=0}^{\infty} P(U_k = l - i | Q_k = i, \lambda_k, \mu_k) E\{Q_n | Q_{k+1} = l, \lambda_{k+1}, \mu_{k+1}\}.
\end{aligned} \tag{3.17}$$

The probabilities in the last line of equation (3.17) will be obtained from the equation below, which calculates the probability distribution of the relevant increments.

$$
\begin{aligned}
C_k(u) &= p[U_k = A_k - S_k = u | Q_k = i, \lambda_k, \mu_k] \\
&= \sum_{s=0}^{\infty} p(A_k - S_k = u | S_k = s, Q_k = i, \lambda_k, \mu_k) p(S_k = s | Q_k = i, \lambda_k, \mu_k) \\
&= \sum_{s=0}^{\infty} p(A_k = u + s | Q_k = i, \lambda_k, \mu_k) p(S_k = s | Q_k = i, \lambda_k, \mu_k) \qquad (3.18)
\end{aligned}
$$

3.6.5 Numerical results

The initial result of fitting a Poisson regression to discharge data reveals that discharge volume at each time epoch, S_n (i.e. the number of discharged patients during time interval $[n, n+1)$, depends upon workload (census) at that time n. Also, there has been significant evidence for the effect of months on the volume of discharge. Therefore, months have been divided into two different categories: 1) "fall and winter months" starting from September up to the end of February and, 2) "Spring and summer months" from March to August.
Moreover, we have blocks of the day which are statistically significant to be included in our models. Hence, our model is written as,

$$
\begin{aligned}
\log[E(Discharged_n)] =& \beta_0 + \beta_1 Census + \beta_2 I_{Block2} + \beta_3 I_{Block3} \\
&+ \ldots + \beta_8 I_{Block8} + \beta_9 I_{fallandWinterMonth}, \qquad (3.19)
\end{aligned}
$$

which indicates that the discharge number at each time point would come from a Poisson distribution with mean parameter $E(Discharged_n)$. The parameter estimates are given in Table (3.4).

Table 3.4: GLM fit (Poisson) to arrival and discharge data

Arrival				Discharge			
Parameters	Estimate	St.error	P-value	Parameters	Estimate	St.error	P-value
Intercept	2.15	0.020	< 2e-16 ***	Intercept	2.09	0.018	< 2e-16 ***
census	-0.001	0.0004	0.00324 **	census	0.02	0.0004	< 2e-16 ***
BoD-2	-0.38	0.019	< 2e-16 ***	BoD-2	-0.37	0.016	< 2e-16 ***
BoD-3	0.13	0.017	1.65e-13 ***	BoD-3	-0.11	0.015	7.63e-13 ***
BoD-4	0.91	0.015	< 2e-16 ***	BoD-4	-0.06	0.015	2.77e-05 ***
BoD-5	0.92	0.015	< 2e-16 ***	BoD-5	0.10	0.013	1.15e-14 ***
BoD-6	0.79	0.015	< 2e-16 ***	BoD-6	0.14	0.013	< 2e-16 ***
BoD-7	0.69	0.016	< 2e-16 ***	BoD-7	0.16	0.013	< 2e-16 ***
BoD-8	0.53	0.016	< 2e-16 ***	BoD-8	0.16	0.012	< 2e-16 ***
Weekday	0.09	0.007	< 2e-16 ***	fallwinter Month	-0.09	0.007	< 2e-16 ***

The same procedure with a slight change in variable selection was applied to the arrival process as well. In modeling arrivals, monthly effects were no longer significant; instead we had

a weekend or weekday effect which was important to be considered and was entered to model as an indicator variable.

$$\begin{aligned}\log[E(Arrival_n)] =& \beta_0 + \beta_1 Census + \beta_2 I_{Block2} + \beta_3 I_{Block3} \\ &+ \ldots + \beta_8 I_{Block8} + \beta_9 I_{WeekendOrWeekday}.\end{aligned} \tag{3.20}$$

We fit these 2 models and determined the corresponding coefficients for each model. At this point the conditional probabilities for any system state with its specific first lag could be automated using equation (3.18) and substituting for it in (3.17).
This model not only enables us to make 24 hour forecasts into the future, but also by having arrival and discharge distributions we can make confidence intervals for our forecasts too.

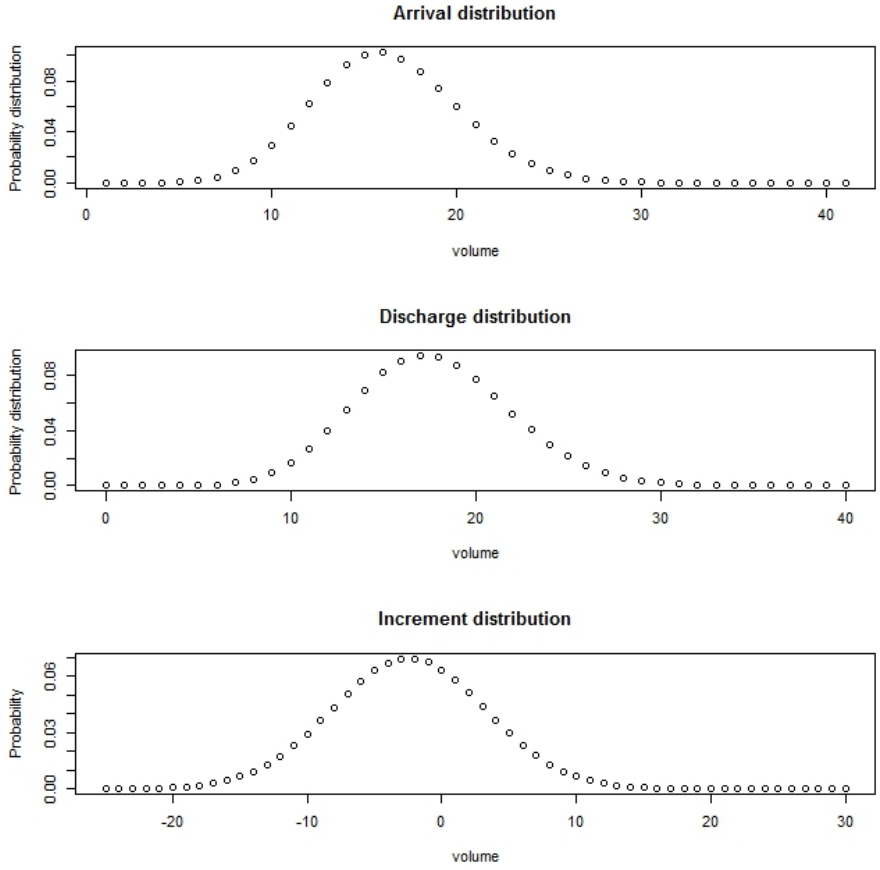

Figure 3.16: Arrival and Discharge distributions.

Graph (3.16) is an illustration of sample arrival, discharge and increment distributions for the 8th block of a week day in a cold month. The first figure in this graph is a prediction for distribution of arrivals; thus, this algorithm not only enables us to have point estimations but also provides us with the the prediction of whole distribution shape. The second figure is a prediction for the discharge distribution while the last one illustrates the increment distribution.

3.7 Application for ED Admins

Knowledge of the future alone cannot solve the overcrowding problem; action based on this knowledge is required too. If any decision for near future interventions is to be made, measuring the efficiency of that decision in terms of how successful it will be in decreasing the backlog, is important too.

Calling in a physician for the entire day or just for a couple of hours when a surge in the census is forecasted, could be among the possible decisions. We present an example of such decisions in each scenario and their effect on the census level.

With our DTMC algorithm, it is possible for ED managers to make short-term predictions of the census at any time during the day. If we assume that at 6:00 am there are 10 patients in the waiting room of an ED, the manager might be interested to know how this number would change in the ED, 1 to 14 hours into the future. Figure (3.17) presents the dynamics of this system.

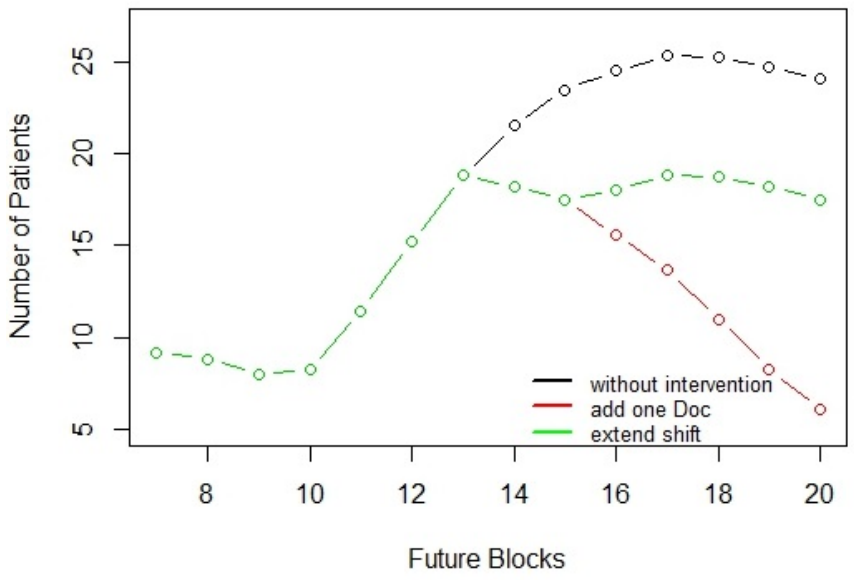

Figure 3.17: ED census predictions.

Meanwhile, if this manager were to decide to add one more physician at time 1:00 pm to avoid the predicted peak in the ED, the red line in the graph shows the effect of her decision. Alternatively, knowing that there will be a shift change at 3:00 pm, she might ask the next shift's physician to arrive 2 hours earlier at time 1:00 to help with the busy ED. The effect of this decision can be seen in the graph with the green line in Figure (3.17).

3.8 Model Validation

At this stage, we wish to see how each of these models performs as compared to other models. In order to validate and compare our models, we have divided our data set into training and test data sets to perform a five-fold cross validation. Since our prediction algorithms depend on previous step(s) to predict the next census, a "sliding-window" approach has been used at each fold.

We use the sliding-window validation technique where a sliding window of training data is used to fit required model and derive parameter estimations [52]. The basis for this approach is to divide the forecast horizon into multiple periods and then to update and extend an existing plan in each period (e.g., Sethi et al. [43]). The number of future periods for which the forecast is made is the *horizon* and these are the periods which *roll over* once a forecast for that period is made (Sethi & Sorger [42]).

This technique ensures that parameters remain up to date throughout the entire forecasting and that there is no single division between the data used for parameter estimation and validation. Once parameters are available, the model will be used to make from 1 up to 8 step forecasts into the future on the entire test data points.

We predict the model parameters from training data and test them on our test data. Commonly used forecast-accuracy metrics which are the Mean Square Error (MSE), and correlation at various forecast lags have been measured and compared for our three models in table (3.5).

Table 3.5: Model Validation

Measures	First Model	Second Model	Third Model
One Step	MSE=32.180 R=0.835	MSE=26.865 R=0.848	MSE=27.234 R=0.848
Two step	MSE=57.436 R=0.703	MSE=39.375 R=0.768	MSE=39.60 R=0.77
Three step	MSE=70.847 R=0.623	MSE=46.130 R=0.722	MSE=47.079 R=0.722
Four step	MSE=75.896 R=0.571	MSE=48.930 R=0.702	MSE=48.03 R= 0.70
Five step	MSE=80.106 R=0.538	MSE=50.753 R=0.690	MSE=51.161 R=0.685

This table explores the reliability of the forecasting methods in term of the correlation coefficient and mean square error measures. The first model (DTMC) performs poorly as compared to other two models as it appears to have the highest MSE and least correlation at each forecasting level. Performance measures decrease as the length of forecasting window increases. For example, the forecasts of the census have correlation coefficients of 0.83, 0.70, 0.62, 0.57 and 0.52 respectively with the actual census count at 1, 2 up to 5 steps (i.e. 3, 6, up to 15 hours) into the future. While both second and third models have the same behaviour in this regard, they perform considerably better than the first model.

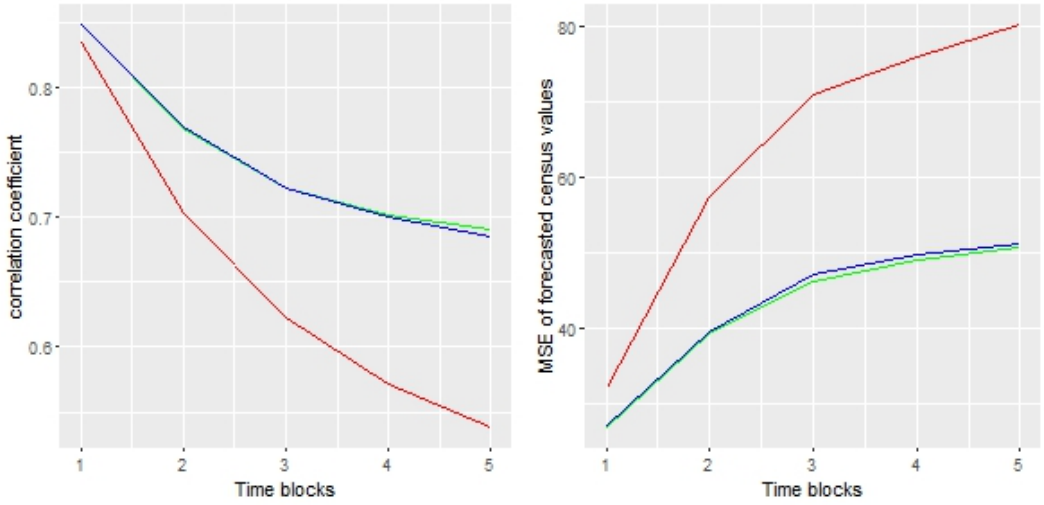

Figure 3.18: Models prediction performance comparison.

It is also important to mention that although the second and third model show almost equal outcome measures in the model validation table, since the third model provides more information on the distribution of forecasted census values, we deem it to be preferable. Having the distribution of census values rather than just a point estimation will help us to find the cumulative distribution functions or the probability of extreme census value occurrences such as sudden surges at any time point into the future.

Figure 3.18 presents a graphical view of the performance of our models. The red, green and blue lines represent the Empirical DTMC model, the regression with ARMA errors and the hybrid model respectively.

3.9 Conclusions

Understanding the dynamics of patient census and accurately predicting future census are essential to the management of staff and resource planning. We presented an approach that not only provides hourly estimations for future census but it also estimates the distribution of the census.

Our Parametric DTMC model could be used in non-stationary systems where there is no exact knowledge of the transient arrival and service rates and they depend on other variables (e.g. calender variables, workload, etc.)

Our findings showed that a parametric discrete time Markov chain algorithm that accounts for recent census levels, arrival and discharge distributions and seasonality, provides higher accuracy of hourly census prediction compared with the empirical model and provides opportunities to reduce biased estimations of patient census. Our results also indicated that the regression model with ARMA errors makes good estimations for the census however, it doesn't provide information about the shape of the distribution.

Our proposed algorithm can also assist decision makers in an emergency department to carry out their own analysis on the impact of their staffing decisions.

Finally, with more data at hand on the number of beds or nursing staff and more information about shifts and physicians, this algorithm could be used to explore various scenarios to address either an impending surge, or an existing backlog, in order to provide timely access to care for the patients.

Chapter 4

The Lowest Priority Waiting Time Distribution in the Affine and the Delayed Accumulating Priority Queues

Maryam Mojalal[1], David A. Stanford[1], Richard J. Caron[2]

1: Department of Statistical & Actuarial Sciences, The University of Western Ontario, London, Canada

2: Department of Mathematics and Statistics, The University of Windsor, Windsor, Canada

4.1 Abstract

The accumulating priority queue (APQ) was first introduced under another name by Kleinrock [28] as a "time-dependent priority queue", where patients accumulate priority as a linear function of the time they spend in the queue. When a server becomes free, the customer with the highest accumulated priority will enter service. The waiting time distribution for each class in the APQ was obtained in Stanford, Taylor, and Ziedins (2014)[10]. Since then, such models have been referred to in the literature as Accumulating Priority Queues (APQ).

All subsequent publications addressing the APQ since then have assumed that all arriving customers accumulate priority credits over time starting from the same initial value (assumed, without loss of generality, to be 0). The model we present herein introduces a new element in terms of an initial class-dependent credit level, from which the accumulated priorities of the various classes grow linearly over time. We consider the case of a two-class APQ, for which class-1 customers receive a positive initial credit upon arrival. The present work is concerned with determining the waiting time distribution for the lower class of customers in such an APQ, and in assessing the impact of the initial priority credit upon that distribution.

As in the health care setting we are concerned about the long waiting times of the lowest acuity

patients, we have addressed that in relation to the initial class-1 credit value, accumulation rate and system occupancy level. Therefore, this work also considers a particular optimization problem related to the model, namely, the selection of the optimum accumulation rate which allows for the lowest class customers to meet their KPIs.

4.2 Introduction

Waiting time distributions for queues operating under a wide variety of service disciplines have been well known for more than 50 years: from the first-come, first-served (FCFS) discipline to the "classical" priority queuing discipline (Kesten and Runnenberg [19]), in which a customer belonging to a given priority class is selected for service only when there are no waiting customers from higher priority classes. Customers from low priority classes in such a situation can be repeatedly overtaken by customers from higher priority classes whenever any are present in the queue. None of these disciplines factor the incurred waiting time in determining a customer's priority. Kleninrock (1964) was the first to introduce a service discipline in which customers earn priority credit as a linear function of their waiting time.

Kleinrock, in his paper, derived a set of recursive formulae to calculate the average waiting time before service for the different classes of patients. However, Stanford et al. [10] derived the Laplace-Stieltjes Transform (LST) of the stationary waiting time distributions for each class of patients in the Poisson arrival, general service and single server case.

Later, Sharif et al. [8] extended the applications of APQ in health, and considered the multi server case with a common exponential service time distribution, and calculated the probabilities of waiting times exceeding specified time limits, which corresponded to "Excess waiting times". The multi service APQ with heterogeneous servers was later studied by Li et al. in 2016 [35], as they believed a model which is capable of modeling heterogeneity in server speed would reflect the reality of health care systems better.

A few years after Kleinrock's initial work, Kleinrock & Finkelstein in 1967 [30] considered power-law functions for priority accumulation. This sort of nonlinear behaviour was revisited by Li et al. (2017)[36] and extended to a larger class of non linear functions. Other APQ papers which have recently appeared include Haviv & Ravner [31], Kella & Ravner [38] and finally Fajardo & Drekic [5] where they investigated different types of preemptive linear APQs.

In all of these studies, the initial priority of patients at the time of entrance is zero while the higher the priority of a patient, the greater the rate at which that patient accumulates priority. However, in this study, we are interested to look at the affine problem where the higher class patient starts initially from a positive credit while the lower class stars with zero credit.

Our reasons for studying the affine variant of the APQ are two-fold. From a theoretical viewpoint, Li et al. [36] recently established the conditions under which a family of non-linear priority accumulation functions are, in fact, completely equivalent to a family of linear priority accumulation functions, in terms of the range of waiting time behaviours that can be obtained. It therefore is of interest to us to explore what the simplest non-linear APQ is which can pro-

duce different behaviours, and that corresponds to the affine APQ model which we will describe fully in the next section.

From a practical viewpoint, our interactions with certain health care professionals have revealed that it does arise in certain settings that low-acuity patients are deemed to be of no particular urgency for treatment until their waiting time reaches a particular time threshold. In an APQ setting, this means that the low-acuity patients are deemed to not accumulate any priority over time until that time threshold has been reached. Even though the theoretical and practical motivations for these two-class models appear to describe differing priority accumulation mechanisms, we establish in the next section that they are, in fact, equivalent.

It follows as a result of the foregoing equivalence that the waiting time distribution of the lower-priority class of customers is identical to those of the lower class in a classical priority queue, up to the time threshold. Beyond that time point, the waiting time behaviour resembles that of a non-affine APQ. We are able to exploit this fact to come up with an algorithm for the determination of the waiting time distribution for those customers who experience waits in excess of the time threshold. Numerical examples will be presented to illustrate the trends we observe.

4.2.1 Description of the Affine & the Delayed Variants of the APQ model

We consider a single-server queue with Poisson arrivals and a common general service time distribution. Customers of class $i; i = 1, 2$ arrive at the queue according to a Poisson process with rate $\lambda_i; i = 1, 2$ respectively. Upon arrival, a customer of class i starts accumulating priority at rate $b_i; i = 1, 2$, where $b_1 > b_2$.

We call this an Affine APQ due to the fact that the higher priority class customers receive an initial priority credit $a > 0$ upon arrival, and therefore accure priority as a function of time t according to the priority function defined by

$$q_1(t) = a + b_1(t - \tau_1), \tag{4.1}$$

where τ_1 denotes the arrival instant of the customer. In contrast,

$$q_2(t) = b_2(t - \tau_2), \tag{4.2}$$

where τ_2 denotes the arrival instant of the class-2 customer. At every service completion instant, the next patient with the highest earned priority at that instant will enter service.

We define $d = a/b_2$ to be the time that it takes for a class-2 customer to accumulate credits equal to a class-1 customers initial credit, a. This provides us with the means to link the affine variant to the delayed variant of the APQ, whose description follows.

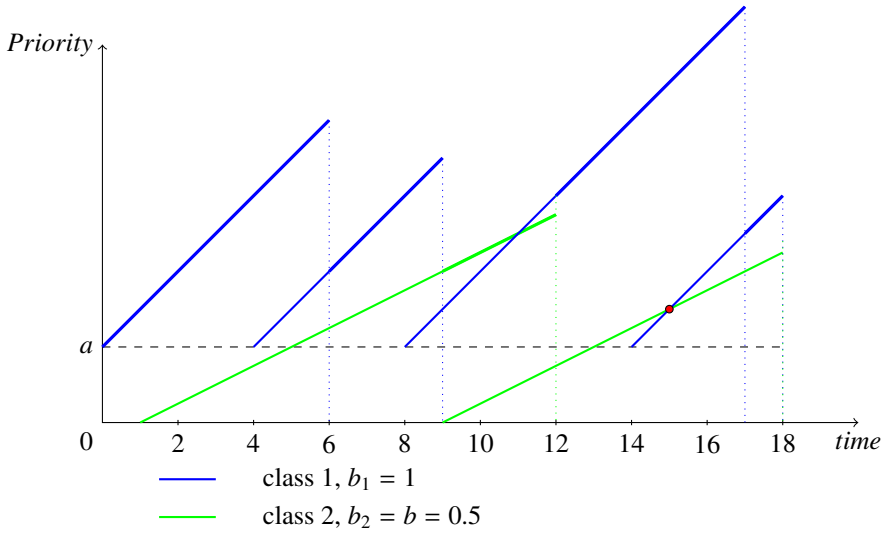

Figure 4.1: Accumulated priorities in an Affine APQ

In the delayed variant of the APQ, class-1 customers receive no priority credit upon arrival, but start accumulating priority immediately at rate b_1, so that $q_1(t) = b_1(t-\tau_1)$ where τ_1 once again refers to the arrival instant of the class-1 customer. It is the class-2 customers who are delayed for a period of time d before they start to accumulate priority. Thus, $q_2(t) = 0; \tau_2 \leq t \leq \tau_2 + d$ while $q_2(t) = b_2(t - d - \tau_2); t \geq \tau_2 + d$.

Figure (4.1) plots the accumulated priorities of customers as a sample path of such processes. Without loss of generality, throughout this study we assume that $b_1 = 1$ and $b_2 = b$ for some $0 \leq b \leq 1$. In Figure (4.1), $b = 0.5$.

We observe that if $b = 0$ a classical priority queue is obtained, while $a = 0$ results in a classical APQ. Also, a classical priority queue would be the result in the limit, if $a \to \infty$.

Similar to the definition 3.1 in Stanford et al. (2014), the Maximum Priority Processes for the Affine APQ are defined as the least upper bounds, respectively, for the accumulated priorities of queued customers from each class at each time t, given only knowledge of the times at which previous customers entered service, and their accumulated priority at these times. Equivalently, a class-1 customer who has accumulated priority more than all class-2 customers currently waiting in the queue is called an accredited customer. An accredited class-1 customer is guaranteed service before any waiting class-2 customer.

An accreditation interval consists of the service time of a non-accredited customer followed by a sequence of service times of accredited class-1 customers. Thus, the busy period of the queue can be broken into a sequence of accreditation intervals.

In Figure (4.1), the red point indicates the time when a higher class customer overtakes a lower class, who had entered the system earlier but will yet go to service after that higher class customer. Stanford et al. (2014) Lemma 4.2 established that during an accreditation interval, the time points at which customers become accredited occur according to a Poisson process with rate $\lambda_1(1 - b_2/b_1)$. By analogy to similar constructs in the classical $M/G/1$ queue, they

were able to obtain the expression of the Laplace-Stieltjes transform (LST) of the duration of an accreditation interval and its mean duration.

In the next section we turn to the determination of the waiting time distribution for the lower class customers in an Affine APQ discipline in the setting as defined above.

4.3 Lower-class Waiting time Distributions in the Affine & Delayed Variants of the APQ under $M/G/1$

The first task we address in this section is to establish the equivalence of the waiting time distributions for the lower priority class in the affine and the delayed variants of the APQ which possess the same priority accrual rates $b_i; i = 1, 2$ and where the affine element $a > 0$ is related to the delay element $d > 0$ by $d = a/b_2$. Having done so, we turn to the determination of this common waiting time distribution for both models, using the affine APQ as the setting for that derivation.

Theorem 4.3.1 *Consider both an affine APQ with parameters* $a > 0, b_1 > 0$ *and* $b_2 > 0$ *and a delayed APQ with the same* b_1 *and* b_2 *as the affine APQ, along with a delay period* $d = a/b_2$ *for class-*2 *customers. Then the stationary waiting time distributions for class-*1 *and class-*2 *customers would be the same in both variants, when subjected to the same traffic load* $\rho = (\lambda_1 + \lambda_2)/\mu < 1$.

Proof Figure 4.2 (A) illustrates the priority accumulation functions that would apply to class-1 and class-2 customers, respectively, both arriving at time 0 in the affine APQ model as stipulated above. Figure 4.2 (B) illustrates the corresponding priority accumulation functions for the delayed variant of the APQ. The dashed line added to 4.2 (B) represents the continuation of the linear portion of the delayed priority function for class-2 customers, back to its y-intercept.

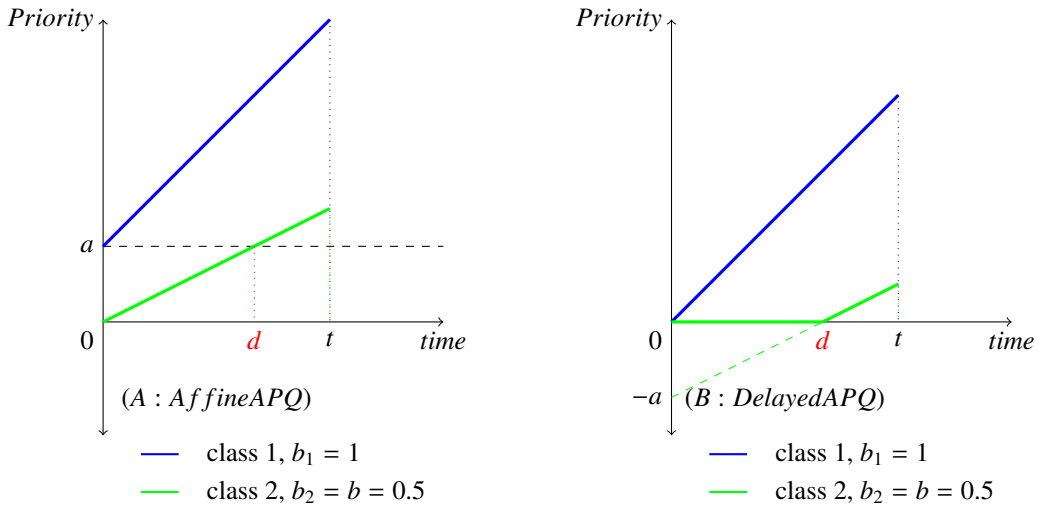

Figure 4.2: Accumulated priorities in an Affine APQ and Delayed APQ

Under the delayed APQ protocol, a class-2 customer would only be served during the first d time units in the event that no class-1 customers were present at such an instant. The same customer selections between class-1 and class-2 customers would be arrived at using the linear function $q_2(t) = -a + b_2 t$, represented by the linear portion and its dashed-line continuation in FIG B.

Now comparing 4.2 (A) and (B), we recognize that the former priority accumulation functions are merely displaced vertically from those in the latter figure, where $q_2(t) = -a + b_2 t$. Thus, the same decisions as to who would be selected for service under either arrangement for any sample path comprising the same arrival instants and sequence of service time durations. It necessarily follows that the waiting time distributions for both classes of customers would be the same in both the affine and delayed APQ variants.

Having established the equivalence of the affine and delayed APQ variants, we turn at this point to the task of determining the waiting time distribution before service commences for class-2 customers. We do so under the framework of the Affine APQ discipline, in terms of its LST. From these results, we will obtain the corresponding distributions under the Classical APQ and classical priority disciplines by appropriate choices of the accumulation rates. In subsequent sections we present some numerical examples which will be obtained via numerical inversion of the corresponding LSTs using the Gaver-Stehfest Algorithm.

From a class-2 customer's perspective, from their arrival moment until time $d = a/b_2$ when this customer earns enough credit to start competing with higher class customers, all higher class customers enter service prior to this customer and this customer experiences a discipline similar to the classical priority queue.

Therefore, the waiting time for a class-2 customer can be viewed as consisting of two cases. The first corresponds to the case where the waiting time is less than time d which will be derived in Lemma 4.3.2. The second addresses the additional waiting time after d which is derived in the sbsequent theorem.

Lemma 4.3.2 *Let $\mathcal{W}_2$ denote the stationary waiting time random variable for class-2 customers in the affine case, and W_2 its counterpart in the standard single-server non-preemptive priority queue with the same arrival rates. Then*

$$P(\mathcal{W}_2 \leq t) = P(W_2 \leq t), \forall t \leq d. \tag{4.3}$$

Proof The key to the proof is to couple the waiting time in the affine APQ with that in the corresponding non-preemptive priority queue over the interval of time $0 \leq t \leq d$.
As argued in the proof of Theorem 9.1 of Stanford et al. (2014), an arriving class-2 customer in the APQ system will only be served once all work found in the system upon arrival has been served as well as the later-arriving accrediting customers. The same is true for a class-2 customer in a standard non-preemptive priority queue, all of whose later class-1 arrivals accredit immediately relative to any class-2 customer present in the system. We therefore exploit the same rearrangement of service times used in the proof of Theorem 9.1 of Stanford et al. (2014) to determine the waiting time distribution in the non-preemptive priority queue:

we first attend to the work present in the system upon the arrival of a tagged class-2 customer, and then turn to the later-arriving class-1 customers.

Since the tagged customer is from a Poisson process at rate λ_2, the workload it perceives upon arrival (which corresponds to time instant $t = 0$) is equal in distribution to the unfinished workload of the server in an $M/G/1$ queue attending to the two customer classes on a FCFS basis. The waiting time for class-2 customers in the standard non-preemptive priority queue is readily obtained by treating this unfinished workload as the initial "delay" as per Conway, Maxwell, and Miller pp 149-151. It then follows that the corresponding waiting time is the duration of the delay cycle comprising the initial "delay" followed by a delay busy period comprising service to the later-arriving class-1 customers, until none are left.

Under the customer-selection rule for the affine APQ, during an interval of time of duration d, all class-1 arrivals subsequent to the arrival of a tagged class-2 customer accredit relative to that customer immediately. Thus, for given realizations of the unfinished workload found upon arrival and sequence of class-1 arrivals during this period of duration d, there is no distinction between who would be served ahead of the tagged customer under the affine APQ discipline and the standard non-preemptive priority discipline.

In other words, for any given pair of realizations such that the unfinished workload and subsequent delay busy period of later-arriving class-1 customers has been completed under the non-preemptive priority discipline at a given point in time $0 \leq t \leq d$, it would likewise end at the same instant $0 \leq t \leq d$ under the affine APQ, and vice versa. This establishes the Lemma above.

The foregoing lemma enables us to invoke the waiting time distribution for class-2 customers when their waiting time is less than d. However, one further concept is needed in order to state the relevant results for when their waiting time exceeds d. Class-2 customers who wait more than d time accumulate priority in excess of a. From this point onward, further accreditations by class-1 customers will be according to a Poisson process at rate $\lambda_1(1 - b_2/b_1)$, as in the case with the non-affine APQ. The LST of the wait time for class-2 customers in this case is derived according to the following theorem.

Theorem 4.3.3 (M/M/1) *Let $\mathcal{W}_{q_2}$ denote the stationary waiting time random variable for a class-2 customer operating under the affine APQ discipline under the foregoing stated assumptions and Exponential service time distribution. Let $\phi(s)$ be the LST of the common exponential service time distribution, and let N_t be the number of customers ahead of the tagged class-2 customer at time t. $I\{A\}$ denotes the indicator function of the event A.*
Without loss of generality, we assume the tagged class-2 customer arrives at time 0. The Laplace transform associated with the stationary distribution of $\mathcal{W}_{q_2} I\{\mathcal{W}_{q_2} > d\}$ is given by

$$E\{e^{-s\mathcal{W}_{q_2}} I\{\mathcal{W}_{q_2} > d\}\} =$$
$$\sum_{i=1}^{\infty} P\{N_0 = i\} \sum_{j=1}^{\infty} P\{N_d = j, N_t > 0; 0 \leq t \leq d | N_0 = i\} e^{-sd} (\eta_1^A(s))^j, \tag{4.4}$$

where $\eta_1^A(s)$ is the LST of the distribution of the busy period length and would be derived as

$$\eta_1^A(s) = \phi(s + \lambda_1(1-b)(1-\eta_1^A(s))). \tag{4.5}$$

Proof The result is obtained by deriving the conditional LST first. Upon arrival, in order to be delayed in excess of d time, the tagged class-2 customer must find some number $i \geq 1$ of customers in the system ahead of them, $N_0 = i$. Furthermore, at no point during the period of time $0 \leq t \leq d$ can N_t drop to zero, or the waiting time of the tagged customer would end at such an instant. After d time elapses, this customer must find some number $j \geq 1$ customers ahead of them, who are either those of i initial customers or any higher class customer who has arrived later and overtaken this customer, since they have higher priority. Hence, the conditional LST is given by

$$\begin{aligned} &E\{e^{-s\mathcal{W}_{q_2}}(I\{\mathcal{W}_{q_2} > d\})|N_0 = i, N_t > 0; 0 \leq t \leq d, N_d = j\} \\ &= e^{-sd}(\eta_1^A(s))^j, \end{aligned} \tag{4.6}$$

where $\eta_1^A(s)$ represents the LST of the distribution of the length of busy periods initiated by each of those j customers and continued by arriving class-1 customers who overtake the tagged customer with rate $\lambda_1(1-b)$. Summing over the possible values of i and j the final result is obtained.

Theorem 4.3.3 above addresses the APQ discipline in the case of a common exponential service time distribution for both classes (i.e. under $M/M/1$ queue).
A similar result can be obtained for more general service times, but to do so, one needs to be aware of Theorem 3.1 in Adan & Haviv (2009) [21], which establishes the nature of the dependence between the number present in queue at the end of a service time and the duration of that service time. (The exponential service case is shown to be an exception in Adan & Haviv (2009) so that the derivation above is correct).

The context is as follows. According to a well known property, in any non-preemptive, work-conserving M/G/1 queue under stationary conditions the residual life time, R, and elapsed lifetime (age), A, of the service length of the customer in service have the same distribution. Furthermore, the residual of the service length of the customer in service, has a density function according to

$$f_R(r) = \frac{1 - F_X(r)}{E(X)}, \tag{4.7}$$

where X is the service time distribution. This property has been called the "random modification" in [61]. More details in this regard could be found in [29] (chapter 5.2 page 172).
However, Adan & Haviv, in their paper, show that this property (4.7) is no longer true if further information such as the number of waiting customers in the queue is available. They then obtain the density function and the LST of the conditional age of service, given the number of

customers present in the system (including the one in the service), $Q^+ = n$; $n \geq 1$, under an $M/G/1$ discipline as

$$f_{A|Q^+=n}(a) = \frac{\rho}{\pi_n} f_A(a)[(1-\rho)\frac{(\lambda a)^{n-1}}{(n-1)!} + \sum_{i=1}^{n} \pi_i \frac{(\lambda a)^{n-i}}{(n-i)!}]e^{-\lambda a},\ a \geq 0, \tag{4.8}$$

where $\pi_i; i = 0, 1, 2, ...$ is the stationary distribution of number in system in an $M/G/1$ queue (see Theorem 3.1 in [21]).

Corollary 4.3.4 (M/G/1) *Consider the assumptions in Theorem 4.3.3. Let R denote the residual of the service length of the customer in service at time d and $\phi_g(\cdot)$ be the LST of the general service distribution. The Laplace transform associated with the distribution of $\mathcal{W}_{q_2}(I\{\mathcal{W}_{q_2} > d\})$ under the affine APQ $M/G/1$ discipline is*

$$\begin{aligned} &E\{e^{-s\mathcal{W}_{q_2}}(I\{\mathcal{W}_{q_2} > d\})\} = \\ &\sum_{i=1}^{\infty} P\{N_0 = i\} \sum_{j=1}^{\infty} P\{N_d = j, N_t > 0; 0 \leq t \leq d | N_0 = i\} e^{-sd} (\eta_r^A(s))(\eta_g^A(s))^{(j-1)}, \end{aligned} \tag{4.9}$$

where

$$\eta_g^A(s) = \phi_g(s + \lambda_1(1-b)(1-\eta_g^A(s))) \tag{4.10}$$

is the LST of the distribution of the busy period generated by each customer waiting in the queue ahead of the tagged customer; and, $\eta_r^A(s)$ is the conditional LST obtained as

$$\eta_r^A(s) = \phi_r(s + \lambda_1(1-b)(1-\eta_g^A(s))). \tag{4.11}$$

$\phi_r(s)$ is the conditional distribution of R given $j \geq 1$ customers in the system. The conditional distribution, $\phi_r(s)$ can be derived according to the Equation (4.8).

Finally, the following steps should be carried out in order to obtain the waiting time distribution according to Equation (4.4): (1) finding the $P\{N_0 = i\}$ for the desired queue. (2) obtaining the transition matrix to obtain the relevant transition probabilities between states, $P\{N_d = j, N_t > 0; 0 \leq t \leq d | N_0 = i\}$. (3) finding $\eta_r^A(s)$ and $\eta_g^A(s)$ under the mentioned assumptions of the model.

4.4 Specific details for the algorithm

In this section, we study special cases in the presented algorithm for the Affine APQ. The first numerical example will be related to an Affine APQ under $M/M/1$ discipline, where class-i customers $i = 1, 2$ enter the system according to Poisson processes with rates λ_i. We establish the algorithm (4.4) in the Theorem 4.3.3 to obtain the waiting time distribution for lower class

customers and present the results in a graph. The second numerical example will study the application of the introduced algorithm in the Affine $M/D/1$ case where the common service distribution for customers is a deterministic distribution.

4.4.1 $M/M/1$

We initially seek to substitute three components in (4.4) with their relative values.

As for the first component $P\{N_0 = i\}$, in a single-server $M/M/1$ queueing system under a non-preemptive and work-conserving service, the steady state distribution is Geometric. The distribution of N_0, being what an arrival from a Poisson process sees, is therefore also Geometric so that $P\{N_0 = i\} = (1-\rho)\rho^i; i = 0, 1, 2, ...$, where $\rho = (\lambda_1 + \lambda_2)/\mu$.

To derive the second component, the taboo conditional transition probability $P\{N_d = j, N_t > 0; 0 \leq t \leq d | N_0 = i\}$, we construct a Continuous Time Markov Chain (CTMC) with the transition rate matrix (infinitesimal generator), Q, for the number of customers ahead of the tagged class-2 customer, N_t, on the interval $0 \leq t \leq d$. This CTMC's transition rate matrix is given by

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0\dots \\ \mu & -(\mu+\lambda_1) & \lambda_1 & 0\dots \\ 0 & \mu & -(\mu+\lambda_1) & \lambda_1\dots \\ 0 & 0 & \mu & -(\mu+\lambda_1)\dots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}. \tag{4.12}$$

In this matrix, state 0 is an absorbing state and should be avoided since it would indicate the case where $\mathcal{W}_{q_2} < d$, which has been dealt with in Lemma 4.3.2.

Hence, the conditional transition probability, $P\{N_d = j, N_t > 0; 0 \leq t \leq d | N_0 = i\}$, is as the $(i, j)th$ element in the transition matrix P. The transition matrix P is obtained after we apply the well known technique of "Uniformization" which approximates a CTMC by assigning a uniformized discrete-time Markov chain to it.

In this approach, the mean time spent in each state in a CTMC, ν_i, is assumed to be the same for all states as $\nu_i = \nu$, $i = 1, 2, ..., n$ where n is the number of states. Therefore, the number of state transitions, $M(d)$, during time d would be according to a Poisson process with rate ν. Therefore,

$$P_{ij}(d) = P\{N_d = j, N_t > 0; 0 \leq t \leq d | N_0 = i\} = \sum_{k=0}^{\infty} P_{ij}^k P(M(d) = k) = \sum_{k=0}^{\infty} P_{ij}^k \frac{e^{-\nu t}(\nu d)^k}{k!} \tag{4.13}$$

where, the equal transitioning rate at each state could be achieved by introducing the $1 - \nu_i/\nu$ fictitious one step transition from states to themselves. Therefore, for $i = j$, $P_{ij} = 1 - \sum_{i \neq j}^{n} q_{ij}/\nu$

and for $i \neq j$, we have q_{ij}/ν; which could be written in matrix form as: $P = I + Q/\nu$ where I is the Identity matrix.

Consequently, the following matrix P would be the corresponding transition matrix for the Q matrix in Equation (4.12).

$$P = \begin{bmatrix} 1 & 0 & 0 & 0\dots \\ q & 0 & p & 0\dots \\ 0 & q & 0 & p\dots \\ 0 & 0 & q & 0\dots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix} \tag{4.14}$$

where $q = \frac{\mu}{\mu+\lambda_1}$ and $p = \frac{\lambda_1}{\mu+\lambda_1}$.

The third component would be the length of the sub-busy period $\eta_1^A(s)$. Since we have assumed that service times have common exponential distribution, the LST of this distribution, $\phi(s)$, is equal to $\frac{\mu}{\mu+s}$. Consequently, according to (4.5) we have:

$$\eta_1^A(s) = \frac{[\mu + s + \lambda_1(1 - b_2/b_1)] - \sqrt{(\mu + s + \lambda_1(1 - b_2/b_1))^2 - 4\lambda_1(1 - b_2/b_1)\mu}}{2(\lambda_1(1 - b_2/b_1))}. \tag{4.15}$$

Having all components defined, we were able to plot the cumulative distribution function for the class-2 customers as in Figures 4.3 when accumulation rate $b = 0.5$ and in Figures 4.4 when $b = 0.8$.

Our graphs illustrate the results for the waiting time distributions of a class-2 customer in an Affine APQ model while comparing 4 types of policies (classical APQ, $d = 6$, $d = 10$, classical priority) for 2 different values of accumulation rate $b = 0.5$ and $b = 0.8$, based on the arrival rates $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ for class-1 and class-2 patients respectively. The service times are exponentially distributed with $\mu = 1$ (resulting in $\rho = 0.8$).

The initial assumptions for $d = 6, 10$ results in the a values to be $3, 5$ when $b = 0.5$ and $4.8, 8$ when $b = 0.8$. In both figures we notice that the affine curves are bounded by the classical priority queue and classical affine queue, indicating that compared to the classical APQ, the positive delay interval, d, is slightly to the benefit of class-1 customers. Also, we notice the cumulative distribution function ($c.d.f.$) of the waiting time for class-2 is stochastically smaller when $d \to \infty$ (i.e., classical priority) and stochastically largest when $d = 0$ (i.e., APQ).

In summary, the graphs illustrate that by introducing the affine element a, we are able to fine-tune the nature of the priority accumulation to give more urgency to higher class-1 patients if required while still valuing lower class patients incurred wait times.

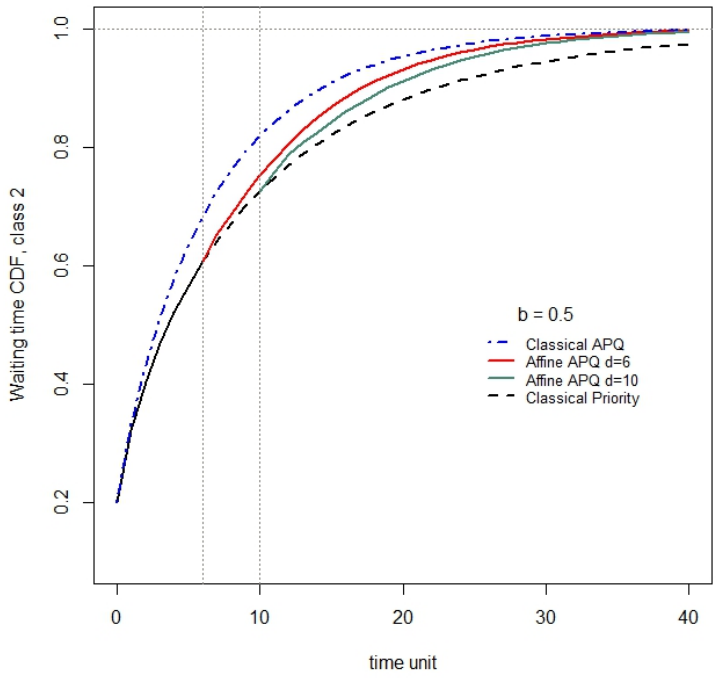

Figure 4.3: GS evaluation of class-2 wait time distribution in affine APQ M/M/1 with $b = 0.5$

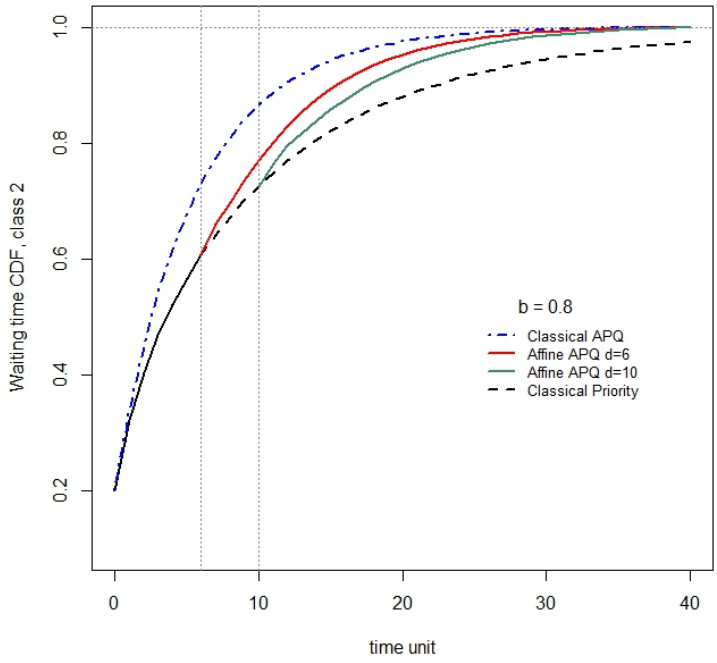

Figure 4.4: GS evaluation of class-2 wait time distribution in affine APQ M/M/1 with $b = 0.8$

4.4.2 $M/D/1$

In the following illustration, we consider an Affine APQ with Poisson arrivals with rates λ_i, $i = 1, 2$ and a Deterministic service time of duration $1/\mu$ for both classes and $c.d.f$ as:

$$F(x) = \begin{cases} 0 & \text{if } x < 1/\mu, \\ 1 & \text{if } x \geq 1/\mu. \end{cases}$$

We recall Equation 2.20 from Chapter 2. Since $M/D/1$ is a special case of the $M/G/1$ queue when all service times are exactly $1/\mu$, the stationary distributions, transition matrix and other relations can be obtained accordingly.

The algorithm represented by Corollary 4.3.4 requires the following four elements to be determined:

1) The initial probability vector: the $M/D/1$ stationary probabilities π_i are given by $\pi_0 = (1-\rho), \pi_1 = (1-\rho)(e^{\rho} - 1)$ and for $i \geq 2$ [14],

$$\pi_i = (1-\rho) \times \{e^{i\rho} + \sum_{k=1}^{i-1} (-1)^{i-k} e^{k\rho} [\frac{(k\rho)^{i-k}}{(i-k)!} + \frac{(k\rho)^{i-k-1}}{(i-k-1)!}]\}. \tag{4.16}$$

2) The conditional transition probability: Substituting the corresponding k_i in the transition matrix for $M/G/1$ queues, we have:

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \dots \\ k_0 & k_1 & k_2 & k_3 \dots \\ 0 & k_0 & k_1 & k_2 \dots \\ 0 & 0 & k_0 & k_1 \dots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

where,

$$k_i = \frac{(\lambda_1/\mu)^i}{i!} e^{-\lambda_1/\mu} = \frac{e^{-\rho}\rho^i}{i!}.$$

3) The length of the sub-busy period generated from the service residual $\phi_r(s)$: Since the service distribution is not exponential and thus not memoryless, the remainder of the service time of a customer who is in the service when the tagged class-2 spends d time in the queue, is no longer exponential. We use the Theorem 3.1 in Adan & Haviv (2009) to derive the distribution as,

$$f_{R|Q^+=n}(r) = \frac{\lambda}{\pi_n} [(1-\rho) \frac{(\lambda(1/\mu - r))^{n-1}}{(n-1)!} + \sum_{i=1}^{n} \pi_i \frac{(\lambda(1/\mu - r))^{n-i}}{(n-i)!}] e^{-\lambda(1/\mu - r)}. \tag{4.17}$$

Hence, the LST of this distribution is given by

$$\begin{aligned}
\phi_r(s) &= \int_{r=0}^{\frac{1}{\mu}} e^{-sr} f_{R|Q^+=n}(r)dr \\
&= \frac{\lambda}{\pi_n} e^{-\lambda/\mu} \Big[\frac{(1-\rho)}{(n-1)!} \lambda^{n-1} \int_{r=0}^{\frac{1}{\mu}} (1/\mu - r)^{n-1} e^{-r(s-\lambda)} dr + \sum_{i=1}^{n} \frac{\pi_i}{(n-i)!} \lambda^{n-i} \\
&\times \int_{r=0}^{\frac{1}{\mu}} (1/\mu - r)^{n-i} e^{-r(s-\lambda)} dr \Big] \\
&= \frac{\lambda}{\pi_n} e^{-\lambda/\mu} e^{-\frac{s-\lambda}{\mu}} \Big[\frac{(1-\rho)}{(\lambda-s)^n} \lambda^{n-1} G_{n,\frac{1}{\lambda-s}}(1/\mu) + \sum_{i=1}^{n} \frac{\pi_i}{(\lambda-s)^{n-i+1}} \lambda^{n-i} \\
&\times G_{n-i+1,\frac{1}{\lambda-s}}(1/\mu) \Big], \qquad (4.18)
\end{aligned}$$

where, $G_{a,b}(.)$ is the CDF for a Gamma distribution with parameters a and b. Therefore, we obtain $\eta_r^A(s)$ according to Equation (4.11).

4) The lengths of the sub-busy periods generated by each of the other customers in the queue $\eta_g^A(s)$: Here we seek to derive the lengths of the sub-busy periods generated during the deterministic service time of the $j-1$ customers waiting in the queue ahead of the tagged class-2 customer after time d. Let $\phi_g(s)$ be the LST of the deterministic service time duration, given by

$$\phi_g(s) = e^{-s(\frac{1}{\mu})}$$

therefore,

$$\eta_g^A(s) = e^{-(s+\lambda_1(1-b_2)(1-\eta_g^A(s)))\frac{1}{\mu}},$$

which should be solved numerically.

4.4.3 $M/M/c$

We turn now to the elements needed to obtain the waiting time distribution for class-2 customers under Affine APQ discipline in a multi-server setting. Class-i; $i = 1, 2$ customers arrive according to Poisson processes with rates $\lambda_i; i = 1, 2$ respectively. The common service distribution is Exponential with rate μ.

We begin by observing that the waiting time in the queue is strictly positive only if an arrival finds all c servers busy, and otherwise it is 0. This probability could be derived from a classical $M/M/c$ queue. Similarly, the number of customers that the tagged class-2 customer finds in the system at the instant of arrival, i, could be identified from the stationary distribution, π_i in an $M/M/c$ queue.

Following the same logic as in the multi-server classical APQ in Lemma 3.1 in [8], we observe that the waiting time LST for customers waiting beyond d period of time, in the Affine APQ with multi-server, can be obtained after some adjustments to the service rate in the single-server according to Corollary 4.4.1. In fact, when all servers are busy, the times between service completions are exponentially distributed with parameter $c\mu$ since there are c exponential servers and each of them are serving at rate μ.

Also during the first d period of time upon arrival, as mentioned in Lemma 4.3.2, class-2 customers see the system similar to a classical priority queue and their waiting time distributions are derived according to the waiting time distributions in a classical priority queue under $M/M/c$ discipline. The following Corollary provides the algorithm to derive the LST of the stationary waiting time distribution under the afore mentioned assumptions.

Corollary 4.4.1 *Consider the affine accumulating priority queue where all classes have common exponentially distributed service times, with common mean* $1/\mu$*. The LST associated to this distribution is* $\phi_c(s) = \frac{c\mu}{(c\mu+s)}$*; then,*

$$E\{e^{-s\mathcal{W}_{q_2}}(I\{\mathcal{W}_{q_2} > d\})\} = \sum_{i=c}^{\infty} P\{N_0 = i\} \sum_{j=c}^{\infty} P\{N_d = j, N_t \geq c; 0 \leq t \leq d | N_0 = i\} \times e^{-sd}(\eta_c^A(s))^{j-(c-1)}, \tag{4.19}$$

where, $\eta_c^A(s)$ *is the LST of the distribution of the busy period length and the solution to the equation,*

$$\eta_c^A(s) = \phi_c(s + \lambda_1(1-b)(1-\eta_c^A(s))). \tag{4.20}$$

We note the similarity to (4.5). The difference here is that we are dealing with c servers in parallel at rate μ each, which is indistinguishable when all servers are busy from a single exponential server working at rate $c\mu$. (In the complementary case where there are fewer than c in system at any point during the first d units of time, the waiting time would terminate at that instant and its distribution would be derived from a classical priority queue result specified in Lemma 4.3.2 above. The probability of being delayed would be obtained using the $M/M/c$ model, but the service times would be exponential at rate $c\mu$.)

Figure 4.5 and 4.6 illustrate the GS evaluation at 8 points for the waiting time distribution of class-2 customers under Affine $M/M/2$ when $\lambda_1 = 0.7$ and $\lambda_2 = 0.9$ are arrival rates and $\mu = 1$ is the common service rate. In this figure we notice the same general pattern as of Figures 4.3 and 4.4. Also, we note that the occupancy level is $\rho = 0.8$ in both cases; however, the cumulative probability values for similar time points are larger when the number of servers has increased.

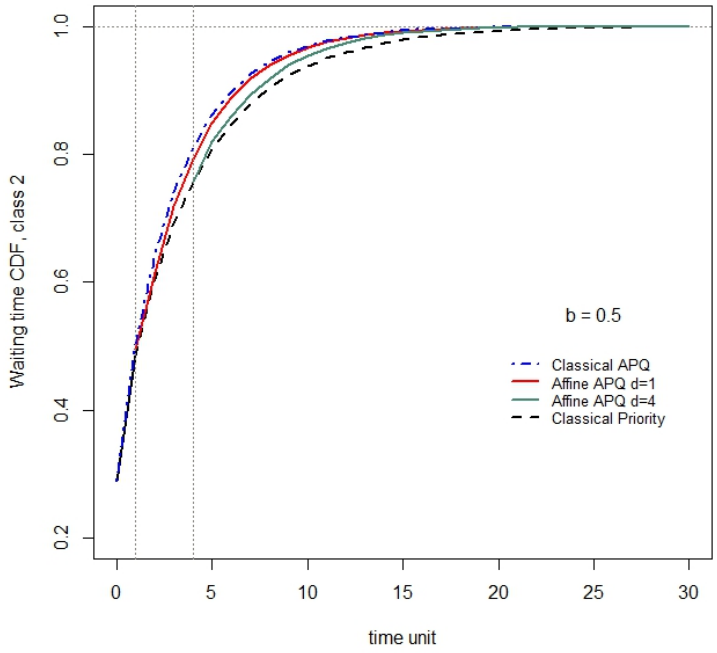

Figure 4.5: The GS evaluation of class-2 waiting time distribution in affine APQ M/M/2

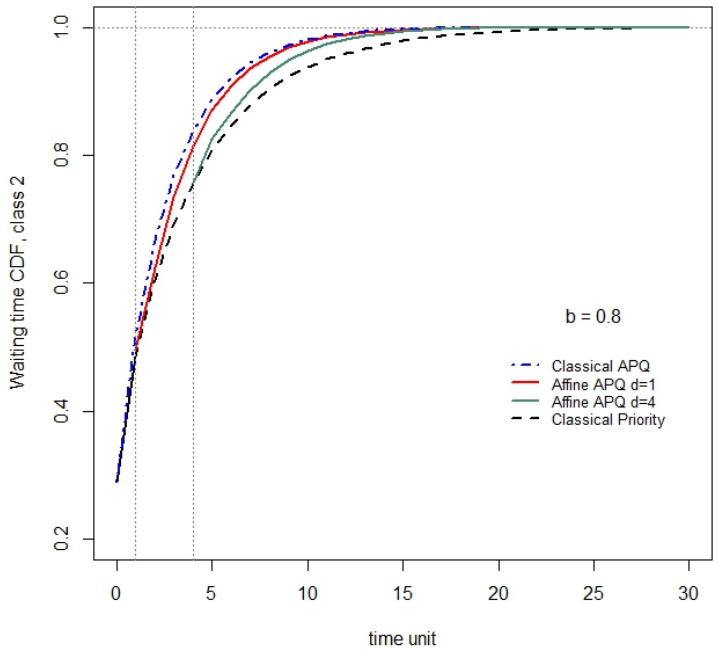

Figure 4.6: The GS evaluation of class-2 waiting time distribution in affine APQ M/M/2

4.5 Numerical investigations to solve for the optimum priority accumulation rate

One potential application of this study could be in health care management after customers are triaged and a decision should be made on which patient to be treated next. A ready example arises in the field of emergency medicine, which serves the needs of patients whose lives are in imminent danger, those of moderate urgency, and others who have comparatively minor complaints.

The Canadian Triage and Acuity Scale, CTAS (2005)[22], classifies patients, based on a primary symptom assessment, to five distinct populations, and sets a service standard for commencement of service for each category (see Table 4.1). These standards specify a delay target and a corresponding compliance probability p for each class, in a way that the probability a patient from the specific class will be seen by a physician before the corresponding delay target is at least p.

Category	Level of acuity	Conditions	Response time	Targets (%)
1	Resuscitation	Threats to life	Immediate	98
2	Emergent	Potential threat to life	15 mins	95
3	Urgent	Progress to a serious problem	30 mins	90
4	Less urgent	Potential for deterioration	60 mins	85
5	Non urgent	May be acute but non-urgent	120 mins	80

Table 4.1: CTAS key performance indicators.

These delay targets are typically set by medical professionals prior to any consideration of the traffic characteristics of the patient classes or the queue. Therefore, the selection of a service discipline such that the delay target needs for each class are accommodated, is within the responsibilities of the health care professionals.

Among the main objectives of studying the waiting times of customers in the APQ's in Stanford et al. [10] which was inspired from health care, was building a system which factors both the time that customers have spent in the system as well as their acuity level (priority class) so as to better adhere to the stated delay targets. In fact, the advantage of an APQ approach for systems operating under KPIs is that the lower priority class customers can still be overtaken by others of greater urgency or acuity, but they will not be overtaken indefinitely, due to the growing accumulated priority the longer a customer waits.

Consequently, Sharif [12] studied an optimisation problem to minimize a weighted average of the total expected excess over all classes of customers. In [8], Sharif et al. aimed to determine a "feasible region" comprising a combination of overall utilisation and specified accumulation rate, for which all classes of customer meet their KPI targets. The authors were able to choose the priority accumulation rates for the various classes, to provide an extra margin of flexibility over the standard priority queueing discipline to ensure the best accumulation rates for different occupancy levels so that both KPIs could be met simultaneously over the widest possible range

of occupancies. In so doing, the best provision is made for possible future growth in demand (and hence a higher occupancy level).

Li (2015) [34] (in Chapter 5) defined the "integrated weighted average excess waiting" (IWAE) function and optimized the IWAE objective in terms of the optimal ratio of the two priority accumulation rates when they focussed on the two-class APQ case .

Having established the waiting time distribution for the lowest priority class in the Affine APQ, here we address the impact of the delay period, d, (equivalently the effect of the initial class-1 credit value, $a > 0$), and the accumulating slope, b, on the chosen delay target KPI for the class-2 waiting time. We will compare systems under different assumptions in terms of the optimum accumulation rate b.

In order to answer this question, we present a series of graphs in different system occupancy levels (ρ), on the x-axis and the values of optimum b on the y-axis, under the assumption of $\lambda_1 = \lambda_2$ and $\mu = 1$. Our goal is to identify a selection scheme and obtain the minimum value of b for which the system would comply with the designated KPI for class-2 customers.

In this study we set the KPI to be such that at least 80 percent of customers should not wait more than 60 minutes in the queue (every unit of time is considered to be 15 minutes; see [17] where Dreyer et al. estimated the mean treatment time for CTAS four patients as 15.0 minutes (95% CI 14.615.4)). Therefore, we formulate this problem as

$$b^* = \min\{b | b \in [0,1]; P(\mathcal{W}_{q_2} \leq 4) \geq 80\%\}, \tag{4.21}$$

where $\mathcal{W}_{q_2}$ is the waiting time distribution for the lower priority class and 4 units of time represent 60 minutes.

We would study this for both one server and two server affine APQ with exponential arrivals and common exponential service in 3 different scenarios: $d = 0, 1, 2$, where $d = 0$ actually represents a classical APQ.

Note that $0 \leq b \leq 1$, as b increases, $\lambda_1(1-b)$ decreased; therefore, class-1 patients will overtake less class-2 patients ahead of them. As a result, the waiting time for lower priority patients will decrease (for fixed values of ρ and d) and the increase in the cumulative probability function at 4 follows naturally. Consequently, in order to find the minimum b we followed these steps:

1. Let $k = 0$, $n = 0$ and $m = 1$ and let $f(b) = P(\mathcal{W}_{q_2} \leq 4)$
2. If $f(0) \geq 0.8$ then $b^* = 0$ else
2.1. $k \leftarrow (n + m)/2$
2.2. $b^* \leftarrow k$
2.3. If $f(k) \geq 0.8$ then $m \leftarrow k$, else if $f(k) < 0.8$ then $n \leftarrow k$
2.4. If $b^* - k < 2^{-20}$ break; else go to 2.1
3. Report b^*.

Figure 4.7 presents a one-server case. We have included Figure 4.8 with the same assumptions except for the number of servers is now $c = 2$, in order to address the effect of number of

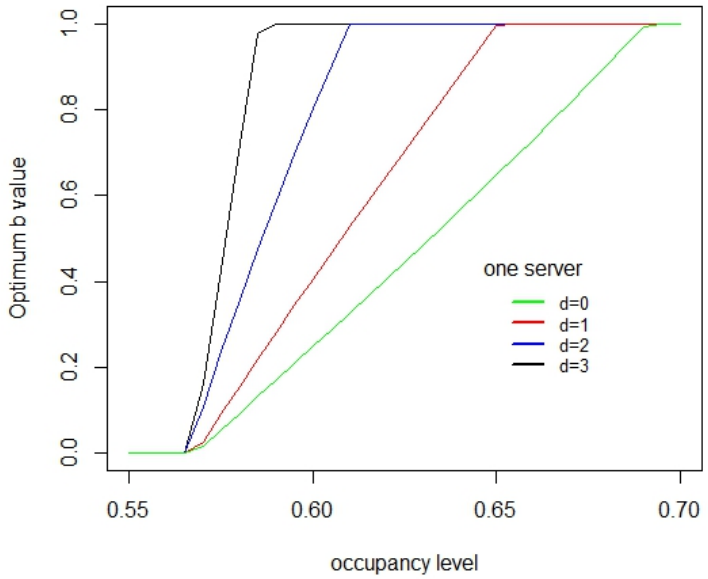

Figure 4.7: Optimum b for different occupancy level, ρ, and d values in Affine APQ under M/M/1

servers. Several facts can be observed from these figures.

1) As the the system's occupancy level (ρ) increases, the value of the optimum b increases so that class-2 customers need to accumulate priority faster in order to meet their target KPI.

2) Class-2 customers can meet their KPI with smaller b values when d is smallest. This result was according to our expectations since when $d = 0$ (the classical APQ), class-2 customers wait less than when under the corresponding affine case.

3) When $\rho \leq 0.552$ even with $b = 0$ (which is equal to the classical priority setting) class-2 customers will be able to meet their KPIs.

4) If we add one more server to the system as in Figure 4.8, for larger occupancy levels class-2 customers can comply with their KPIs with $b = 0$ as compared with the one server case.

In summary, depending upon the traffic patterns at play, it is sometimes possible to find combinations of a and b_2 so that the class-2 KPI is met, without impacting significantly upon the class-1 KPI adherence.

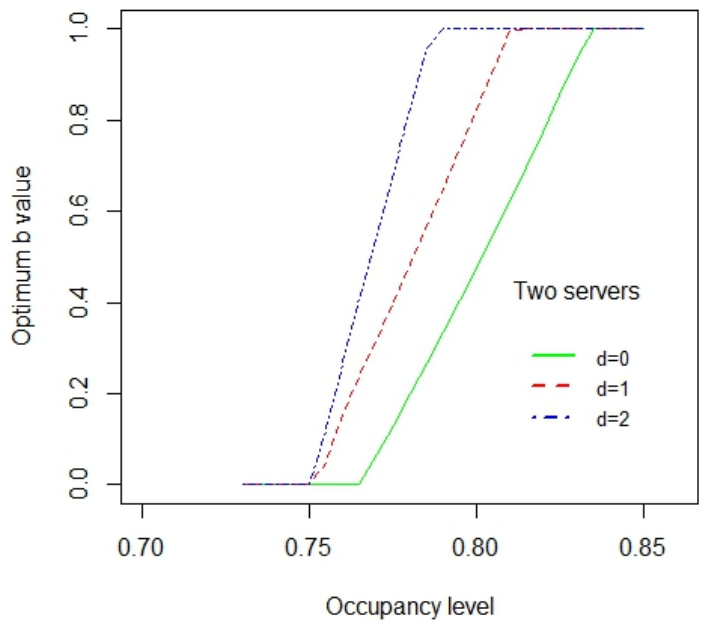

Figure 4.8: Optimum b for different occupancy level, ρ, and d values in Affine APQ under M/M/2

4.6 Conclusion and future research

We initially introduced the Affine APQ and delayed APQ and addressed the importance of the latter disciplines in health care setting. Then, we presented an algorithm to derive the waiting time distribution for the lowest class customers in a one-server, two-class Affine APQ model for which all service times are selected from a common general distribution. This model could easily be extended to include an arbitrary number of classes. We also derived the waiting time distribution for the lowest class customer in an $M/M/c$ Affine APQ discipline.

We obtained the waiting time distributions for some special cases (i.e. M/M/1, M/D/1 and M/M/2) under the Affine APQ discipline and compared their distribution probabilities with the classical APQ and Classical static priority queues.

The present affine APQ models, both single-server and multi-server, enable one to ascertain if a given accumulation rate for lowest class customers will enable compliance with a given KPI for a given traffic pattern of patient arrival and service rate. Therefore, under a pre-assumed delay target KPI for lowest priority class inspired by CTAS KPI's we studied an optimisation problem. Specifically, our objective was to answer the question of what values of class-2 accumulation rate, b, and utilization factor, ρ, enable class-2 customers to meet their assigned KPIs.

Note that this algorithm cannot be used to obtain the waiting time distributions for the higher

acuity patients (class-1 customers in this study). This question could be approached from the following perspective: while a class-2 perceives its situation for the first d period of time (upon its arrival) as comprising a priority queue, a class-1 does not perceive it the same way.

In contrast, upon the arrival instant of a class-1 customer, all of the the class-1 customers present will have to complete service, plus some number (between none and all) of the class-2 customers present who possess more than "a" credits at the arrival instant. By the time that all of the class-1 customers present have been served, it may well be the case that the new class-1 customer will have overtaken some of the class-2s who were present with in excess of a credits at that arrival instant. In this way, the affine variant of the APQ serves to regulate class-1 waiting times better, in that the longer it waits, the less the work there is likely to be ahead of it.

As a result, waiting time distributions for class-1 customers have a more complicated structure and require more information about the maximum credit of class-2 customers. This inspires our next research in the following chapter, where we study the stationary bi-variate distribution of the maximum priority processes of class-1 and and class-2 at the instant of service commencement. Having the exact value of the earned credit by a customer who commences a service will give the waiting time of that customer after a simple re-scale.

Chapter 5

The bivariate Maximum Priority Process in an Affine APQ

Maryam Mojalal[1], Na Li, Peter Taylor[2], David A. Stanford[1]

1: Department of Statistical & Actuarial Sciences, The University of Western Ontario, London, Canada

2: Mathematics and Statistics, The University of Melbourne, Melbourne, Australia

5.1 Abstract

We study the Analysis of the Maximum Priority Processes in the context of Affine APQ. We derive the LST of the stationary steady state distributions of the Maximum Priority Processes as recursive functions in the Affine APQ setting and obtain the explicit solutions for the LSTs in the classical APQ. We employ this argument to present a new approach to determine the LST of waiting time distribution for an APQ with two-classes of customers under the $M/M/1$ discipline. Since the Analysis of the Maximum priority processes in this work is done for the general class of Affine APQs, it provides the grounds for future research to obtain the LST of the waiting time distributions in Affine APQs.

5.2 Introduction

In 1964, Kleinrock [28] proposed a time-dependent priority queueing model where the rate of accumulating priority reflects customers' urgency (classification). In this model, waiting customers accumulate priority as a linear function of the time they spend in the queue. He derived a set of expressions for the expected waiting times for different classes, under the assumptions of Poisson arrivals and a single server working at an exponential service rate.

Stanford et al. [10] obtained the stationary waiting time distribution for each class of customers using the idea of Maximum Priority Processes. The bivariate Maximum Priority Process is a stochastic process which provides the least upper bound for the priorities of customers from each class. After introducing the concept of the accreditation intervals they were able to express the busy period into the sum of accreditation intervals and derive expressions for the actual credit of the customers in each class when they enter into service. Finally, they presented expressions for the Laplace transform of the waiting time distributions in each class. Since then, such models have been referred to in the literature as Accumulating Priority Queues (APQ).

Another method to study the waiting times in an APQ was presented by Fajardo et al. [4], where they introduced the "q-policy" control queues and applied it to get the same results as in Stanford et al. (2015).

In the previous chapter (see Chapter 4), we presented an algorithm to obtain the waiting time distribution for the lower class customers in an Affine Accumulating Priority Process (APQ) setting under M/G/1 and M/M/c disciplines. While this algorithm was applied to derive class-2 customers' waiting times in some special examples with two classes of customers (see Section 4.4), it couldn't be used to obtain corresponding distributions for class-1 customers. The underlying reason is that upon arrival, a class-1 customer may see a group of class one and two customers in front of them. Within classes the First Come First Served (FCFS) discipline applies; therefore, all class-1 customers ahead of this tagged customer will be served first. Whereas, the situation is unclear about those class-2 customers who had waited long enough and gained credit more than a prior to the arrival of this tagged customer. The hidden benefit for class-1 customers is that since their accumulating rate is higher, the longer they wait for service they overtake more class-2 customers. So, the longer they wait, the fewer people they have to wait for.

Stanford et al. (2014) [10] established the nature of the equivalence between the accumulated priority at the time of service commencement for each customer and their waiting time in a non-affine APQ. Consequently, the LST for the stationary accumulated priorities at the time points that customers enter service also can be used to obtain the LST for the stationary waiting times by a re-scaling of the arguments. This important fact motivated us to analyse the bivariate Maximum Priority Processes in an Affine APQ. For some preliminary details about Affine APQs see Subsection 2.4.2.

As the value of the Maximum Priority Process at the instant a new service commences reveals the exact credit value of the customer entering into the service, we aim to derive the Laplace transforms for the steady state joint distributions of $M_1(t)$ and $M_2(t)$, in an Affine APQ, at the mentioned instants. We present a set of recursive equations, and give explicit solution for the LSTs when $a = 0$. We discuss how we can employ these results in order to derive the Laplace transform for the waiting time distributions and present the required theory and expressions.

The outline of this chapter is as follows: We introduce the possible states of the bivariate Maximum Priority Process at service commencement instants in an Affine APQ in section 5.3. We obtain the transition densities (kernels) from and to each set of states in subsection 5.3.2.

Section 5.4 presents the LST of of the joint stationary distribution of $M_1(t)$ and $M_2(t)$. In the last section we present the required relations to derive the LST of the waiting time distributions for both priority classes in an APQ.

Since the transition probability densities of the bivariate Maximum Priority Processes are obtained for the Affine APQs, it can open the grounds for future research to obtain the LST of the waiting time distributions in the Affine APQs as well.

5.3 Limiting distributions in an Affine APQ

We are interested in deriving the limiting distributions for the bivariate Maximum Priority Process (defined in Definition 2.4.1 in Chapter 2) in the Affine Accumulating Priority Queues (Affine APQ) under $M/M/1$. This bivariate process at immediately after a service completion instant (or the moment a new service commences) forms a Markov process with a continuous state space.

If our bivariate process were a Markov chain defined on a discrete state space, then we could employ Theorem 4.3.3 (page 175) in [32] to obtain the limiting distribution, $\pi = [\pi_0, \pi_1, ...]$. For any irreducible ergodic discrete-time Markov chain $\lim_{n\to\infty} P^n_{ij} = \pi_i$ exists and is independent of i. Furthermore, π_i is the unique nonnegative solution of $\pi = \pi P$ where P is the transition matrix. In this case π is called the stationary distribution.

However, the bivariate MP process is a Markov process. For a Markov process (on a continuous state space), the probability distribution π is called **the stationary distribution** if

$$\pi(y) = \int p(x, y)\pi(x)dx, \qquad y \in \mathcal{S}; \tag{5.1}$$

or equivalently,

$$\pi(A) = \int p(x, A)\pi(x)dx, \tag{5.2}$$

for all measurable sets $A \subset \mathcal{S}$, where $\mathcal{S}$ denotes the state space in a Markov process. $p(x, y)$ is called the **transition density** (or the **transition kernel**).

The Markov process is governed by a transition kernel $p(x, A)$ for $x \in \mathcal{S}$ and $A \subset \mathcal{S}$. The main requirement for this Markov process to reach its stationary distribution is that it is irreducible and aperiodic. The irreducibility is defined as, for any $x, y \in \mathcal{S}$, there always exists a positive integer n such that $p^n(x, y) > 0$. In other words, the Markov process can jump into any state from any other state in a finite number of steps. The aperiodicity means that there exist no subsets of the state space $\mathcal{S}$ that can only be periodically visited by the Markov process [44].

Having established all required background, now as the first step to obtain the stationary distributions we identify all possible state sets. Later on, for this embedded Markov process at the beginning of service instant, we develop equations to obtain the stationary distributions.

5.3.1 Identification of possible states

If we plot the states of the system at the end of service events, there are five types of states for the bivariate Maximum Priority Process as shown in Figures 5.1, 5.2 and 5.3 (where the two graphs in the Figure 5.2 are grouped as one in the analysis throughout this chapter).

Suppose that at time t_0 a customer arrives to an empty system, and this customer's service time ends at time $t > t_0$. Then the five possible types of states for the bivariate Maximum Priority Process at time t can be identified as follows:

1) an empty system, for which both MP processes reach their lower limits $(a, 0)$, as in Figure 5.1 (A).
2) a system with no class-1 customers present at an end of service, but a class-2 customer with some amount $y < a$ units of credit present, as in Figure 5.1 (B).
3) The two maximum priority processes are equal, with level (x, x) where $x > a$; as in Figure 5.2 (both situations).
4) Both processes are above their respective minima such that $M_2(t) = y < M_1(t) = x$. The situation where $M_2(t) > a$ at the service completion instant is depicted in Figure 5.3 (A), while $M_2(t) < a$ in Figure 5.3 (B).

In the first state Figure 5.1 (A), immediately after a service completion $M_1(t) = a$ and $M_2(t) = 0$. This situation implies that there is no customer in the system and the idle period will start until a customer enters in the system. The second state (Figure 5.1 (B)) refers to all situations where immediately after a service completion (or at the instant a new service commences), a class-2 customer with accumulated credit less than a is the one who starts the service. therefore, this scenario refers to $M_2(t) < a$ while $M_1(t) = a$.

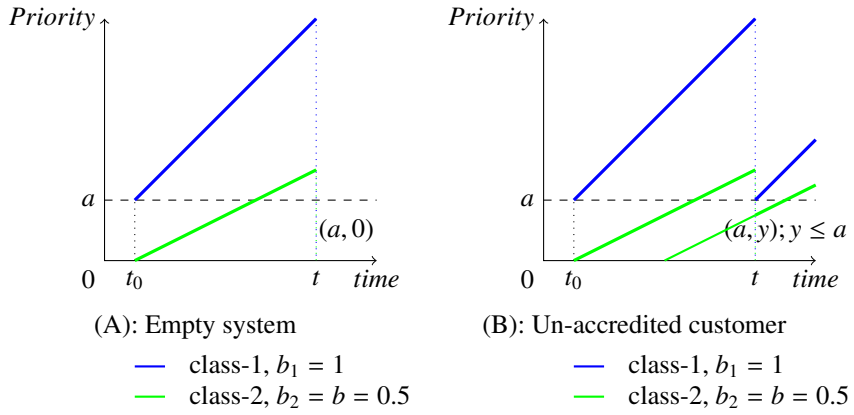

Figure 5.1: $(a, 0)$ and (a, y) states in an Affine APQ

The next set of possible states (5.2) occur when at the absence of any accredited customer, a

non-accredited customer enters the service immediately after a service completion. This non accredited customer could be a class-1 customer whose accumulated priority is less than $M_2(t)$, where obviously $M_2(t) > a$, or a class two customer who has been in the queue long enough to gain credit more than a.

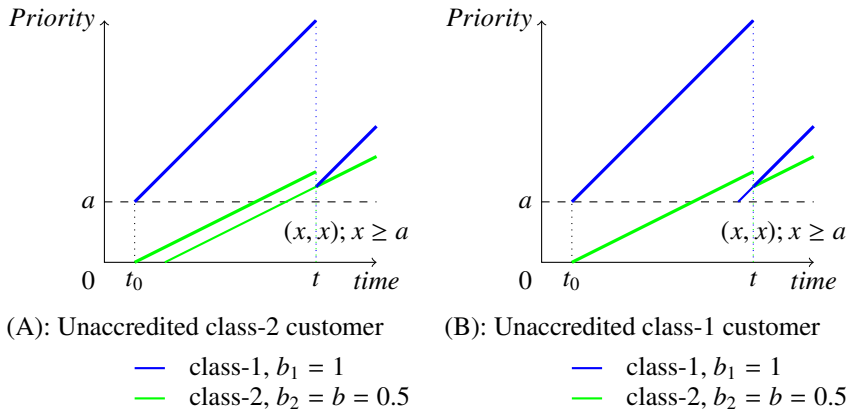

Figure 5.2: An unaccredited customer enters service (state (x, x))

This state by itself, as illustrated in Figure 5.2, could happen in two different scenarios. The first case, which occurs when $M_2(t)$ is less than a, indicates at least one accredited class-1 customer is ready to go into service where as a need to know basis we are only aware of the $M_2(t) < a$. The second case also refers to an at least one accredited customer in the queue while $M_2(t) > a$.

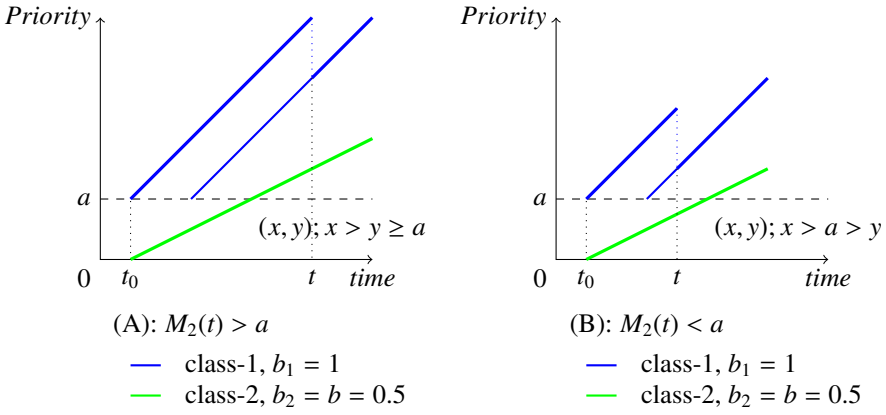

Figure 5.3: An accredited class-1 customer enters service (state (x, y))

For illustration purpose only, in all of the figures above $b_1 = 1$ and $b_2 = b = 0.5$. In the following section we will be deriving transition densities from and to each set of states.

5.3.2 Derivation of transition densities

In this subsection, we specify the transition kernel densities of transitioning from one state set at the start of service event instant t_0, to the other state set at the start of the next service

event.

When the process is transitioning to an empty system (so that $M_1(t) = a$ and $M_2(t) = 0$ until the next arrival instant), the transition occurs with a discrete probability, denoted by $P_{x\to y}$. For example $P_{(a,0)\to(a,0)}$, corresponds to the case where both service completions leave the system empty. In contrast, all other transitions have corresponding probability density function values denoted by expressions of the form $f_{A\to B}$.

Now, let $\lambda = \lambda_1 + \lambda_2$ refer to the total arrival rate of the Poisson arrival process, and let the exponentially-distributed service times be at rate μ. Class-2's accumulation rate is $b_2 = b$ and for class-1 it is $b_1 = 1$.

1. Transition from state $(a, 0)$

We start with the probability of transitioning from $(a, 0)$ to the next such occurrence. This corresponds to the situation where no arrivals occur during the service time, u, of the customer who started the service; thus,

$$\begin{aligned} P_{(a,0)\to(a,0)} &= \int_0^\infty \mu e^{-\mu u} e^{-(\lambda_1+\lambda_2)u} du \\ &= \frac{\mu}{\mu+\lambda} \\ &= \frac{1}{1+\rho}. \end{aligned} \tag{5.3}$$

To briefly describe the particulars of Equation (5.3), a service begins after an idle period and lasts for u units of time, the conditional probability of no arrival during that service is $e^{-(\lambda_1+\lambda_2)u}$, while $\mu e^{-\mu u}$ is the relative likelihood of a service time of length u.

In like manner, the bivariate Maximum Priority Process can reach $(a, y); y \le a$ from an empty system in case of no class-1 arrivals during the service time u, but at least one class-2 arrival such that the $M_2(t)$ drops to y at the end of service time. Note that starting from 0, at the end of service time of duration u, $M_2(t) = bu$. Thus there must be a drop in $M_2(t)$ from bu to y. The likelihood of this event is written as

$$\begin{aligned} f_{\{(a,0)\to(a,y),y\le a\}} &= \int_{\frac{y}{b}}^\infty \mu e^{-\mu u} e^{-\lambda_1 u} \frac{\lambda_2}{b} e^{-\frac{\lambda_2}{b}(bu-y)} du \\ &= \frac{\mu}{\mu+\lambda}\frac{\lambda_2}{b} e^{-\frac{\mu+\lambda_1}{b}y}. \end{aligned} \tag{5.4}$$

In this expression and the following ones, the term $\mu e^{-\mu u}$ refers to the length of service U which is exponentially distributed at rate μ. $M_2(t)$'s drop over an interval is also exponentially distributed with rate $\frac{\lambda_2}{b}$ and an $M_1(t)$ drop is likewise exponentially distributed with rate $\frac{\lambda_1}{b_1} = \lambda_1$ as $b_1 = 1$ by initial assumption.

The event of transitioning from an empty system to state $(x, x); x \geq a$ happens when there is no accredited class-1 customer by the end of service (i.e. there is no customer in the interval $(M_2(t), M_1(t))$). On the other hand, there is at least one customer present of some type, whose credit lies between a and $M_2(t)$. Refer to Figure 5.2 in this regard. Therefore the likelihood of this transition is written as

$$\begin{aligned} f_{\{(a,0)\to(x,x),x\geq a\}} &= \int_{\frac{x}{b}}^{\infty} \mu e^{-\mu u} e^{-\lambda_1(u+a-bu)} (\lambda_1 + \frac{\lambda_2}{b}) e^{-(\lambda_1+\frac{\lambda_2}{b})(bu-x)} du \\ &= \frac{\mu}{\mu+\lambda} (\lambda_1 + \frac{\lambda_2}{b}) e^{-\frac{\mu+\lambda_1^A}{b}x - \lambda_1 a}, \end{aligned} \tag{5.5}$$

where the notation $\lambda_1^A = \lambda_1(1-b)$ is used to denote the rate at which class-1 customers accredit relative to class-2 when $M_2(t) > a$. Starting from a, $M_1(t)$ grows to $a + u$ at the end of service time of duration u. There should be no class-1 arrival in an interval of length $a + u - bu$ (which explains the part $e^{-\lambda_1(u+a-bu)}$ in the above expression) and a drop of both $M_1(t)$ and $M_2(t)$ processes from bu to x which happens according to an exponential distribution with rate $\lambda_1 + \frac{\lambda_2}{b}$.

Finally, for both cases in Figure 5.3 we write;

$$\begin{aligned} f_{\{(a,0)\to(x,y)\}} &= \frac{\mu}{b} e^{-\mu\frac{y}{b}} \lambda_1 e^{-\lambda_1(\frac{y}{b}+a-x)} \\ &= \frac{\mu}{b} \lambda_1 e^{-(\mu+\lambda_1)\frac{y}{b}} e^{\lambda_1(x-a)}; \quad x > a, x > y. \end{aligned} \tag{5.6}$$

In this case the length of service is fixed at $u = y/b$, where $Y \sim Exp(\mu/b)$. Also, at least one class-1 accredited customer should be such that $M_1(t)$ drops from $y/b+a$ to x. This drop occurs according to an exponential distribution with rate λ_1.

Since the transition kernel (density) is a probability function, it has been verified that all probabilities add up to one. (see (A.1) in Appendix A.2.1)

Figure 5.4 demonstrates the admissible region over which $M_1(t)$ and $M_2(t)$ could take values with positive likelihoods. The heavy black lines are the boundaries of the region. There is a point mass probability at $(a, 0)$, the vertical line connecting $(a, 0)$ to (a, a) represents $(a, y), y \leq a$ and the points along the line $y = x$ are corresponding to the (x, x) situation. The line $y = b(x - a)$ indicates the fact that during a service of length u, $M_1(t)$ is bounded by $u + a$ or $\frac{y}{b} + a$ from above.

2. Transition from state set $(a, y); y \leq a$

In this part, we derive transition densities from a point in state set $(a, y); y \leq a$ to other state sets. Applying the same analogy as with the previous case, we derive these transition probabilities accordingly. We start by transitioning to an empty system in one step:

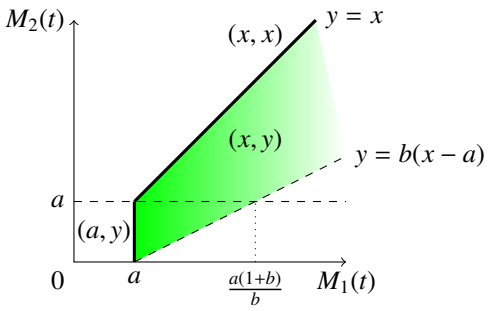

Figure 5.4: Admissible priority regions in an Affine APQ.

$$f_{\{(a,w)\to(a,0);w\le a\}} = \int_0^\infty \mu e^{-\mu u} e^{-\lambda_1 u} e^{-\lambda_2(u+\frac{w}{b})} du$$
$$= \frac{\mu}{\mu+\lambda} e^{-\frac{\lambda_2}{b}w}. \tag{5.7}$$

This situation requires that there be no class-1 arrival during the service time of duration u so that $M_1(t)$ will drop to a. Likewise, there can be no class-2 arrival during the service time of duration u plus the wait time of the class-2 customer who started the service, $\frac{w}{b}$. Under these specifications the busy period will end and system will go idle.

The following equation concerns the density of moving from (a, w) to (a, y). Depending upon the final position of $M_2(t)$ at the end of each service termination, we face two different scenarios: $w \le y$ or $w > y$. If $w > y$ we write:

$$f_{\{(a,w)\to(a,y),w\le y\le a\}} = \int_{\frac{y-w}{b}}^\infty \mu e^{-\mu u} e^{-\lambda_1 u} \frac{\lambda_2}{b} e^{-\frac{\lambda_2}{b}(w+bu-y)} du$$
$$= \frac{\mu}{\mu+\lambda} \frac{\lambda_2}{b} e^{-\frac{\mu+\lambda_1}{b}(y-w)}. \tag{5.8}$$

Figure 5.5 elaborates on this matter. In fact, in the event $w \le y$, the service time must be sufficiently long to enable $M_2(t)$ to grow from w to y; which is why we have $\frac{y-w}{b}$ for the lower bound of the integral.

$$f_{\{(a,w)\to(a,y),y\le w\le a\}} = \int_0^\infty \mu e^{-\mu u} e^{-\lambda_1 u} \frac{\lambda_2}{b} e^{-\frac{\lambda_2}{b}(w+bu-y)} du$$
$$= \frac{\mu}{\mu+\lambda} \frac{\lambda_2}{b} e^{-\frac{\lambda_2}{b}(w-y)}. \tag{5.9}$$

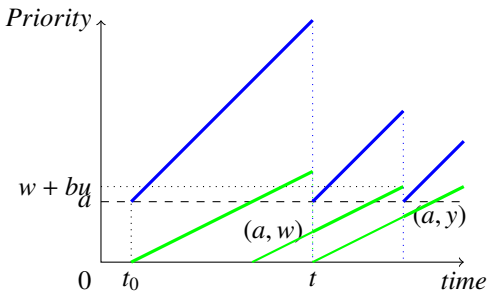

Figure 5.5: Transition $(a, w) \to (a, y)$ when the duration of service time is u and $w \le y \le a$

In moving from (a, w) to (x, x), there should be no customers present at time t accumulated credit between $(M_2(t), M_1(t))$, while the maximum credit between a and $M_2(t)$ is x; thus, there would be a drop with rate $\lambda_1 + \frac{\lambda_2}{b}$ over $bu + w - x$.

$$\begin{aligned} f_{\{(a,w)\to(x,x),x\ge a\}} &= \int_{\frac{x-w}{b}}^{\infty} \mu e^{-\mu u} e^{-\lambda_1(u+a-(bu+w))}(\lambda_1 + \frac{\lambda_2}{b}) e^{-(\lambda_1+\frac{\lambda_2}{b})(bu+w-x)} du \\ &= \frac{\mu}{\mu+\lambda}(\lambda_1 + \frac{\lambda_2}{b}) e^{-\lambda_1 a-(\mu+\lambda_1)\frac{w}{b}+(\mu+\lambda_1^A)\frac{x}{b}}. \end{aligned} \tag{5.10}$$

In like fashion we can write,

$$f_{\{(a,w)\to(x,y)\}} = \frac{\mu}{b} e^{-\frac{\mu}{b}(y-w)} \lambda_1 e^{-\lambda_1(a+\frac{y-w}{b}-x)}. \tag{5.11}$$

The fact that $(M_1(t), M_2(t) = (a, w)$ at the starting time point, affects the possible values they can take at the next time point. Therefore the admissible region in this case is more limited compared to Figure 5.6.

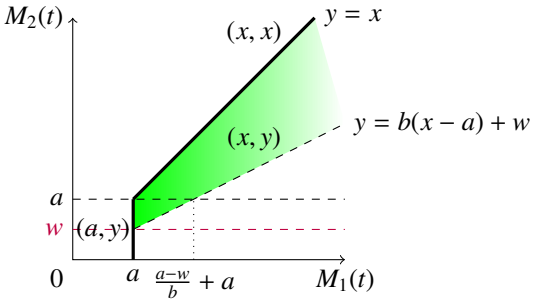

Figure 5.6: Accumulated priorities in an Affine APQ; (a, y)

See Appendix A.2.1 Equation (A.2) for verifications of the pdf assumption.

3. Transition from state set (ν, ν)

Following the same assumptions as before, we seek to obtain one-step transition probabilities from (ν, ν) to other state sets.

$$\begin{aligned} P_{\{(\nu,\nu)\to(a,0);\nu>a\}} &= \int_0^\infty \mu e^{-\mu u} e^{-\lambda_1(u+\nu-a)} e^{-\lambda_2(u+\frac{\nu}{b})} du \\ &= \frac{\mu}{\mu+\lambda} e^{-[(\lambda_1+\frac{\lambda_2}{b})\nu-\lambda_1 a]}. \end{aligned} \tag{5.12}$$

The scenario above requires that there be no class-1 arrival during the service time u plus no waiting class-1 customers with credit less than ν, and no class-2 arrival during the service time u plus the waiting time of the previous class-2 customer $\frac{\nu}{b}$.

$$\begin{aligned} f_{\{(\nu,\nu)\to(a,y);y\le a\le\nu\}} &= \int_0^\infty \mu e^{-\mu u} e^{-\lambda_1(u+\nu-a)} \frac{\lambda_2}{b} e^{-\frac{\lambda_2}{b}(\nu+bu-y)} du \\ &= \frac{\mu}{\mu+\lambda}\frac{\lambda_2}{b} e^{-[\lambda_1(\nu-a)+\frac{\lambda_2}{b}(\nu-y)]}. \end{aligned} \tag{5.13}$$

In the transition from $(\nu, \nu) \to (x, x)$ we must distinguish whether ν is less or larger than x. If $x > \nu$ then the service time should neseccariy be larger than $\frac{x-\nu}{b}$ to provide enough time for extra credit accumulation.

$$\begin{aligned} f_{\{(\nu,\nu)\to(x,x);a\le\nu\le x\}} &= \int_{\frac{x-\nu}{b}}^\infty \mu e^{-\mu u} e^{-\lambda_1(u+\nu-(bu+\nu))}(\lambda_1+\frac{\lambda_2}{b}) e^{-(\lambda_1+\frac{\lambda_2}{b})(\nu+bu-x)} du \\ &= \frac{\mu}{\mu+\lambda}(\lambda_1+\frac{\lambda_2}{b}) e^{-(\mu+\lambda_1^A)\frac{x-\nu}{b}}, \end{aligned} \tag{5.14}$$

$$\begin{aligned} f_{\{(\nu,\nu)\to(x,x);a<x<\nu\}} &= \int_0^\infty \mu e^{-\mu u} e^{-\lambda_1(u+\nu-(bu+\nu))}(\lambda_1+\frac{\lambda_2}{b}) e^{-(\lambda_1+\frac{\lambda_2}{b})(\nu+bu-x)} du \\ &= \frac{\mu}{\mu+\lambda}(\lambda_1+\frac{\lambda_2}{b}) e^{-(\lambda_1+\frac{\lambda_2}{b})(\nu-x)}. \end{aligned} \tag{5.15}$$

In the next case, both x and y should be larger than ν which is larger than a. Also since x is bounded by $u + \nu$ we can write $x < u + \nu$ where $u = \frac{y-\nu}{b}$. Figure 5.7 illustrates this.

$$f_{\{(\nu,\nu)\to(x,y);a<\nu<y<x\}} = \frac{\mu}{b} e^{-\mu\frac{y-\nu}{b}} \lambda_1 e^{-\lambda_1(\frac{y-\nu}{b}+\nu-x)} \tag{5.16}$$

.

See Appendix A.2.1 Equation (A.3) for verifications of the pdf assumption.

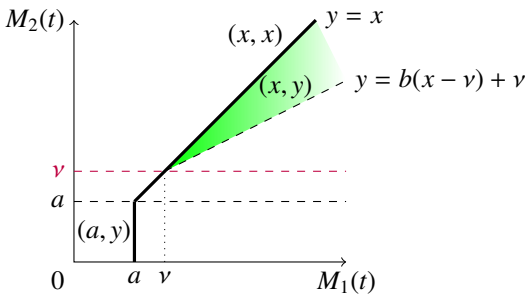

Figure 5.7: Accumulated priorities in an Affine APQ; (x, x)

4. Transition from state set (ν, w)

The last series of transitions are from state set (ν, w). As indicated earlier, the position of w with respect to a makes a difference in the analysis. The first expression pertains to the probability of moving to an idle system. This expression is valid for all values of w (i.e. $w \geq a$ and $w < a$).

$$\begin{aligned} P_{\{(\nu,w)\to(a,0)} &= \int_0^\infty \mu e^{-\mu u} e^{-\lambda_1(u+\nu-a)} e^{-\lambda_2(u+\frac{w}{b})} du \\ &= \frac{\mu}{\mu+\lambda} e^{-\lambda_1(\nu-a)-\frac{\lambda_2}{b}w}. \end{aligned} \tag{5.17}$$

In both cases there should be no arrivals during the service time and no waiting customer at the instant that the service began.

To derive $f_{\{(\nu,w)\to(a,y);y<a\}}$ when $a < w$ we obtain,

$$\begin{aligned} f_{\{(\nu,w)\to(a,y);y<a<w\}} &= \int_0^\infty \mu e^{-\mu u} e^{-\lambda_1(u+\nu-a)} \frac{\lambda_2}{b} e^{-\frac{\lambda_2}{b}(w+bu-y)} du \\ &= \frac{\mu}{\mu+\lambda} \frac{\lambda_2}{b} e^{-\lambda_1(\nu-a)-\frac{\lambda_2}{b}(w-y)}. \end{aligned} \tag{5.18}$$

However, when $w < a$; there can be two possibilities, $w < y < a$ or $y < w < a$. Thus, we have:

$$\begin{aligned} f_{\{(\nu,w)\to(a,y);w<y<a\}} &= \int_{\frac{y-w}{b}}^\infty \mu e^{-\mu u} e^{-\lambda_1(u+\nu-a)} \frac{\lambda_2}{b} e^{-\frac{\lambda_2}{b}(w+bu-y)} du \\ &= \frac{\mu}{\mu+\lambda} \frac{\lambda_2}{b} e^{-\lambda_1(\nu-a)-(\mu+\lambda_1)\frac{y-w}{b}}. \end{aligned} \tag{5.19}$$

Recall that in such a situation, enough time is required so that w increases to reach y, u (or the length of service) starts at $\frac{y-w}{b}$.

$$f_{\{(v,w)\to(a,y);y<w<a\}} = f_{\{(v,w)\to(a,y);y<a<w\}} \tag{5.20}$$

Now we are interested in $f_{\{(v,w)\to(x,x);a<w\le x\}}$; If $w > a$ we have the following two scenarios:

$$\begin{aligned} f_{\{(v,w)\to(x,x);a<w\le x\}} &= \int_{\frac{x-w}{b}}^{\infty} \mu e^{-\mu u} e^{-\lambda_1(u+v-(bu+w))}(\lambda_1 + \frac{\lambda_2}{b}) e^{-(\lambda_1+\frac{\lambda_2}{b})(w+bu-x)} du \\ &= \frac{\mu}{\mu+\lambda}(\lambda_1 + \frac{\lambda_2}{b}) e^{-\lambda_1(v-w)-(\mu+\lambda_1^A)\frac{x-w}{b}}, \end{aligned} \tag{5.21}$$

$$\begin{aligned} f_{\{(v,w)\to(x,x);a<x<w\}} &= \int_{0}^{\infty} \mu e^{-\mu u} e^{-\lambda_1(u+v-(bu+w))}(\lambda_1 + \frac{\lambda_2}{b}) e^{-(\lambda_1+\frac{\lambda_2}{b})(w+bu-x)} du \\ &= \frac{\mu}{\mu+\lambda}(\lambda_1 + \frac{\lambda_2}{b}) e^{-\lambda_1(v-w)-(\lambda_1+\frac{\lambda_2}{b})(w-x)}. \end{aligned} \tag{5.22}$$

However, if $w < a$ then we have

$$f_{\{(v,w)\to(x,x);w<a\le x\}} = f_{\{(v,w)\to(x,x);a<w\le x\}}. \tag{5.23}$$

Finally, we have the following derivation which applies to all ranges of $w < y$.

$$f_{\{(v,w)\to(x,y);w<y<x\}} = \frac{\mu}{b} e^{-\mu\frac{y-w}{b}} \lambda_1 e^{-\lambda_1(\frac{y-w}{b}+v-x)}. \tag{5.24}$$

The admissible regions in these cases would be according to Figures 5.8 and 5.9.

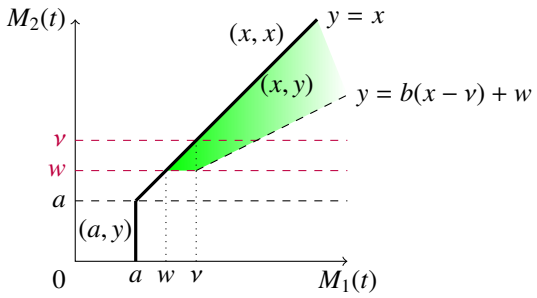

Figure 5.8: Accumulated priorities in an Affine APQ; $(v, w; w > a)$

Having derived all transition densities, in the next section we obtain the stationary distributions related to each state set according to equation (5.2).

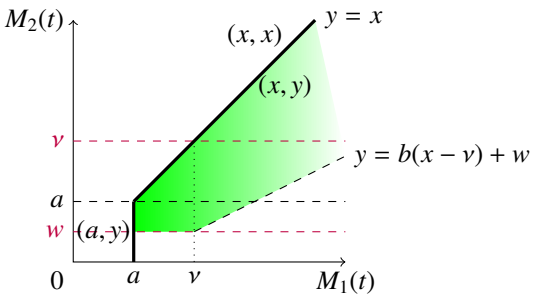

Figure 5.9: Accumulated priorities in an Affine APQ; $(v, w; w < a)$

5.4 Derivation of the LSTs of the limiting distributions

In this section, we first establish recursive equations to derive the stationary distributions of the bi-variate Markov process $(M_1(t), M_2(t))$. In order to explicitly derive these distributions we will proceed to obtain their Laplace transforms.

We should recall Equation (5.2) in light of which the stationary distribution for each state set is obtained. Note that in the following derivations the integrals depend upon the permissible ranges of $M_1(t)$ and $M_2(t)$.

$$\pi(a,0) = \pi(a,0)P_{(a,0)\to(a,0)} + \int_0^a \pi(a,w)P_{\{(a,w)\to(a,0);w\le a\}}dw + \int_a^\infty \pi(v,v)P_{\{(v,v)\to(a,0);v>a\}}dv$$
$$+ \int_0^a \int_a^{\frac{w}{b}+a} \pi(v,w)P_{(v,w)\to(a,0)}dvdw + \int_a^\infty \int_w^{\frac{w}{b}+a} \pi(v,w)P_{(v,w)\to(a,0)}dvdw. \tag{5.25}$$

Next, we obtain the stationary distribution for $\pi(a, y)$. Recall Equations (5.19) and (5.20) where two different density functions are applied in moving from $(v, w); w < a$ to $(a, y); y < a$ depending upon the final level y with respect to w. Also, note that the lowest $M_1(t)$ possible is a, therefore the lowest level it can drop to is a.

$$\pi(a,y) = \pi(a,0)f_{\{(a,0)\to(a,y);y\le a\}} + \int_0^y \pi(a,w)f_{\{(a,w)\to(a,y);w\le y\le a\}}dw + \int_y^a \pi(a,w)$$
$$\times f_{\{(a,w)\to(a,y);y\le w\le a\}}dw + \int_a^\infty \pi(v,v)f_{\{(v,v)\to(a,y)\}}dv + \int_a^\infty \int_w^{\frac{w}{b}+a} \pi(v,w)$$
$$\times f_{(v,w)\to(a,y);y<a<w}dvdw + \int_0^y \int_a^{\frac{w}{b}+a} \pi(v,w)f_{\{(v,w)\to(a,y);w<y<a\}}dvdw$$
$$+ \int_y^a \int_a^{\frac{w}{b}+a} \pi(v,w)f_{\{(v,w)\to(a,y);y<w<a\}}dvdw. \tag{5.26}$$

For the same reason as above turning to $\pi(x, x)$ we obtain,

$$\pi(x,x) = \pi(a,0) f_{\{(a,0)\to(x,x);a\le x\}} + \int_0^a \pi(a,w) f_{\{(a,w)\to(x,x);a\le x\}} dw + \int_a^x \pi(\nu,\nu) \times f_{\{(\nu,\nu)\to(x,x);a\le\nu\le x\}} d\nu + \int_x^\infty \pi(\nu,\nu) f_{\{(\nu,\nu)\to(x,x);a<x<\nu\}} dw + \int_a^x \int_w^{\frac{w}{b}+a} \times \pi(\nu,w) f_{\{(\nu,w)\to(x,x);a<w\le x\}} d\nu dw + \int_x^\infty \int_w^{\frac{w}{b}+a} \pi(\nu,w) f_{(\nu,w)\to(x,x);a<x<w} d\nu dw + \int_0^a \int_a^{\frac{w}{b}+a} \pi(\nu,w) f_{\{(\nu,w)\to(x,x);w<a\le x\}} d\nu dw. \tag{5.27}$$

Finally, depending whether $M_2(t) = y$ is larger or smaller than a, there are two different sets of states. Therefore, distinguishing the cases $y < a$ or $y \ge a$ we obtain:

$$\pi(x,y;y<a) = \pi(a,0) f_{\{(a,0)\to(x,y)\}} + \int_0^a \pi(a,w) f_{\{(a,w)\to(x,y)\}} dw + \int_0^y \int_a^{\frac{w}{b}+a} \pi(\nu,w) \times f_{\{(\nu,w)\to(x,y);w<y<a\}} d\nu dw. \tag{5.28}$$

Note that less terms are involved in Equation (5.28) since fewer transitions are possible which end in state set $\pi(x,y;y<a)$. In face, this event relates to situations in which an accredited customer is entering service. Since $M_1(t) = x > M_2(t) = y$ at such instance no drop has occurred in $M_2(t)$ and as such $M_2(t)$ has continued to grow since the last start of the service. Therefore if y is less than a it follows that w must have been less than a as well. Furthermore, in a situation when a customer with $M_1(t) = M_2(t) = \nu > a$ starts service, for the same reason just discussed, the process can never move to a point where $M_2(t) < a$.

Finally, the limiting distribution for the only remaining state set is obtained as follows,

$$\pi(x,y;y\ge a) = \pi(a,0) f_{\{(a,0)\to(x,y)\}} + \int_0^a \pi(a,w) f_{\{(a,w)\to(x,y)\}} dw + \int_a^\infty \pi(\nu,\nu) f_{\{(\nu,\nu)\to(x,y)\}} d\nu + \int_0^a \int_a^{\frac{w}{b}+a} \pi(\nu,w) f_{\{(\nu,w)\to(x,y);w<a<y\}} d\nu dw + \int_a^y \int_w^{\frac{w}{b}+a} \pi(\nu,w) f_{\{(\nu,w)\to(x,y);a<w<y\}} d\nu dw. \tag{5.29}$$

At this stage, we define Laplace transforms of stationary distributions for each state set as follows,

1. $\tilde{\pi}_Y(s) = \mathcal{L}_Y\{\pi(a,y)\} = \int_0^a e^{-sy}\pi(a,y)dy$
2. $\tilde{\pi}_X(s) = \mathcal{L}_X\{\pi(x,x)\} = \int_a^\infty e^{-sx}\pi(x,x)dx$
3. $\tilde{\pi}_{X,Y;Y\le a}(s_1,s_2) = \int_0^a e^{-s_2 y}\int_a^{\frac{y}{b}+a} e^{-s_1 x}\pi(x,y)dxdy$
4. $\tilde{\pi}_{X,Y;Y>a}(s_1,s_2) = \int_a^\infty e^{-s_2 y}\int_y^{\frac{y}{b}+a} e^{-s_1 x}\pi(x,y)dxdy$

5. $\tilde{\pi}_{X,Y}(s_1, s_2) = \tilde{\pi}_{X,Y;Y\leq a}(s_1, s_2) + \tilde{\pi}_{X,Y;Y>a}(s_1, s_2)$

The above definitions help us simplify the limiting distributions. The more detailed steps have been skipped from this section and transferred into the Appendix. See Equations (A.8) to (A.24) whenever required for more information.

We have a reference point in the following Lemma, which will be useful in all cases.

Lemma 5.4.1 *In an Affine APQ setting under M/G/1, with the Affine parameter a (i.e. the initial class-*1 *credit at the time of entrance into the system), the limiting distribution at the point* $(a, 0)$*, (idle system), is* $\pi(a, 0) = 1 - \rho$*.*

Proof The arrivals from both classes occur according to Poisson processes. From the PASTA property we know that such arrivals see time averages upon arrival. Therefore the stationary probability that a customer arrives to an idle queue is $1 - \rho$ when $\rho = (\lambda_1 + \lambda + 2)/\mu$. Likewise, $\pi(a, 0)$ represents the proportion of time the system is idle.

After evaluating the integrals in Equation (5.25), we obtain the following equation linking the probability of a completely empty system and several of the transforms defined above, evaluated at the specified values:

$$\rho\pi(a, 0) = \tilde{\pi}_Y(\frac{\lambda_2}{b}) + e^{\lambda_1 a}\tilde{\pi}_X(\lambda_1 + \frac{\lambda_2}{b}) + e^{\lambda_1 a}\tilde{\pi}_{X,Y}(\lambda_1, \frac{\lambda_2}{b}). \tag{5.30}$$

In a series of following lemmas, we obtain the LST for the remaining state sets.

Lemma 5.4.2 *The LST of the stationary distribution for state set* $(a, y); y < a$ *is according to:*

$$\begin{aligned}\tilde{\pi}_Y(s) = B(s)^{-1}\Big[&\frac{1 - e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s + \frac{\mu+\lambda_1}{b}}(1 - \rho) + \frac{1 - e^{-(s-\frac{\lambda_2}{b})a}}{s - \frac{\lambda_2}{b}}[\rho(1 - \rho)] + \frac{e^{\lambda_1 a}}{s - \frac{\lambda_2}{b}} \\ &\times \Big(e^{-(s-\frac{\lambda_2}{b})a}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1, \frac{\lambda_2}{b}) - \frac{\mu + \lambda}{b(s + \frac{\mu+\lambda_1}{b})}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1, s)\Big) \\ &- \frac{e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s + \frac{\mu+\lambda_1}{b}}e^{\lambda_1 a}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1, -\frac{\mu + \lambda_1}{b})\Big],\end{aligned} \tag{5.31}$$

when $B(s) = \frac{b(1+\rho)}{\lambda_2} - \frac{1-e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s+\frac{\mu+\lambda_1}{b}} + \frac{1-e^{-(s-\frac{\lambda_2}{b})a}}{s-\frac{\lambda_2}{b}}$.

Proof See Appendix A.2.1, Section A.2.

Lemma 5.4.3 *The LST of the stationary distribution for state set* $(x, x); x \geq a$ *is according to:*

$$\tilde{\pi}_X(s) = C^{-1}\Big[\frac{e^{-(s+\frac{\mu+\lambda_1^A}{b})a}}{s+\frac{\mu+\lambda_1^A}{b}}\Big(e^{-\lambda_1 a}(\pi(a,0)+\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b}))+\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big)$$
$$+\Big(\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}-\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big)\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s-\lambda_1)$$
$$+\frac{e^{-(s-(\lambda_1+\frac{\lambda_2}{b}))a}}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big(\tilde{\pi}_X(\lambda_1+\frac{\lambda_2}{b})+\tilde{\pi}_{X,Y;Y>a}(\lambda_1,\frac{\lambda_2}{b})\Big)\Big], \tag{5.32}$$

where $C(s) = \frac{b(1+\rho)}{\lambda_1 b+\lambda_2} - \frac{1}{s+\frac{\mu+\lambda_1^A}{b}} + \frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}$.

Proof See Appendix A.2.1, Section A.2.

Lemma 5.4.4 *The LST of the stationary distribution for state set* $(x,y); y < a$ *is according to:*

$$\tilde{\pi}_{X,Y;Y\le a}(s_1,s_2) = \frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}e^{-as_1}\Big[\frac{1-e^{-a(s_2+\frac{\mu+s_1}{b})}}{(s_2+\frac{\mu+s_1}{b})}-\frac{1-e^{-a(s_2+\frac{\mu+\lambda_1}{b})}}{(s_2+\frac{\mu+\lambda_1}{b})}\Big]$$
$$\times((1-\rho)+\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b}))+\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}e^{-a(s_1-\lambda_1)}\Big[\frac{e^{-a(s_2+\frac{\mu+s_1}{b})}}{-(s_2+\frac{\mu+s_1}{b})}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})$$
$$+\frac{1}{(s_2+\frac{\mu+s_1}{b})}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s_2+\frac{s_1-\lambda_1}{b})+\frac{e^{-a(s_2+\frac{\mu+\lambda_1}{b})}}{(s_2+\frac{\mu+\lambda_1}{b})}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})$$
$$-\frac{1}{(s_2+\frac{\mu+\lambda_1}{b})}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s_2)\Big]. \tag{5.33}$$

Similarly for the state set $(x,y); y \ge a$ *we will have:*

$$\tilde{\pi}_{X,Y;Y>a}(s_1,s_2) = \frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}\Big[e^{-a(s_1-\lambda_1)}(\frac{e^{-a(s_2+\frac{\mu+s_1}{b})}}{(s_2+\frac{\mu+s_1}{b})})-(\frac{e^{-a(s_1+s_2+\frac{\mu+\lambda_1^A}{b})}}{(s_1+s_2+\frac{\mu+\lambda_1^A}{b})})\Big]$$
$$\times\Big(e^{-\lambda_1 a}(\pi(a,0)+\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b}))+\tilde{\pi}_X(-\frac{\mu+\lambda_1^A}{b})+\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big)$$
$$+\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}\Big[\frac{e^{-a(s_1-\lambda_1)}}{(s_2+\frac{\mu+s_1}{b})}\tilde{\pi}_{X,Y;Y>a}(\lambda_1,\frac{s_1-\lambda_1}{b}+s_2)-\frac{1}{(s_1+s_2+\frac{\mu+\lambda_1^A}{b})}$$
$$\times\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s_1+s_2-\lambda_1)\Big]. \tag{5.34}$$

Proof See Appendix A.2.1, Section A.2.

Equations in (5.31) to (5.34) present a series of recursive functions in terms of Laplace transformations of the stationary distributions for the Maximum Priority Processes in an Affine APQ

(under Poisson arrivals and service times). The explicit solutions to these equations would naturally lead to the LST's of the waiting time distributions for high and low classes of customers under the assumptions of the model.

However, due to the complexity of the equations, we have decided to start with the determination of a complete algorithm for the computation of the bivariate process in the classical, non-affine case where $a = 0$. In this situation, several terms disappear, such as the $\pi(a, y)$ density function; on the other hand, there is more that can be said for the probability masses associated with the integrals over the applicable ranges for the $\pi(x, x)$ and $\pi(x, y)$ terms. In the following section we study the system under this new assumption and find the explicit relations for the LST of the stationary distributions.

5.5 Waiting time distributions when $a = 0$: Classical APQ

Letting $a = 0$, we identify three possible state sets for the Maximum priority processes: state $(0, 0)$ which refers to the instant when our system is idle; state set (x, x) when there is no accredited customer to enter the service at the instant of service completion such that the next customer entering the service is a non-accredited one; and, finally, state set (x, y) when there is at least one accredited class-1 customer in the system to start the next service.

Similar to Figure 5.4 in the previous section, the admissible region for a classical APQ is as indicated in Figure 5.10.

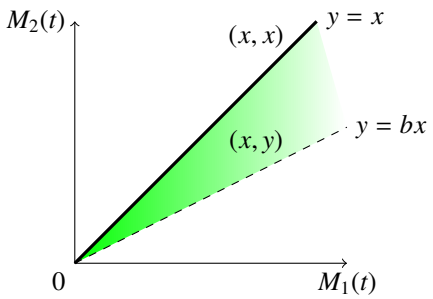

Figure 5.10: Admissible priority region in a classical APQ; ($a = 0$).

Consequently, letting $a = 0$ in the general Equations from (5.31) to (5.34), we observe that $\tilde{\pi}_Y(s)$ disappears, as does $\tilde{\pi}_{X,Y;Y<a}(s_1, s_2)$. The three remaining quantities are evaluated in what follows. First, when $a = 0$ is substituted into Equation (5.30) one obtains:

$$\rho\pi(0, 0) = \tilde{\pi}_X(\lambda_1 + \frac{\lambda_2}{b}) + \tilde{\pi}_{X,Y}(\lambda_1, \frac{\lambda_2}{b}). \tag{5.35}$$

For the next state set (x, x), after letting $a = 0$ and some simplifications as shown in Appendix A.2.1 (see Equation (A.25)) the Laplace transform is written as follows.

$$\tilde{\pi}_X(s) = \frac{(1-\rho)(s-(\lambda_1+\frac{\lambda_2}{b})) - (\frac{\lambda+\mu}{b})\tilde{\pi}_{X,Y}(\lambda_1, s-\lambda_1) + \rho(1-\rho)(s+\frac{\mu+\lambda_1^A}{b})}{\frac{(s+\frac{\mu+\lambda_1^A}{b})(s-(\lambda_1+\frac{\lambda_2}{b}))(1+\rho)}{\lambda_1+\frac{\lambda_2}{b}} - (s-(\lambda_1+\frac{\lambda_2}{b})) + (s+\frac{\mu+\lambda_1^A}{b})}. \tag{5.36}$$

Substitution of (5.36) in (5.35) results in

$$\tilde{\pi}_X(-\frac{\mu+\lambda_1^A}{b}) = -(1-\rho) - (\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b})). \tag{5.37}$$

Finally, for the the state set (x, y), after letting $a = 0$ in equation (5.34) and substituting it in equation (5.37), we obtain the following result:

$$\begin{aligned}\tilde{\pi}_{X,Y}(s_1, s_2) = {} & \frac{\mu}{b}\frac{\lambda_1}{\lambda_1 - s_1}\Big[(\frac{1}{(s_2+\frac{\mu+s_1}{b})}) - (\frac{1}{(s_1+s_2+\frac{\mu+\lambda_1^A}{b})})\Big]\Big(-(\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}))\Big) \\ & + \frac{\mu}{b}\frac{\lambda_1}{\lambda_1 - s_1}\Big[\frac{1}{(s_2+\frac{\mu+s_1}{b})}\tilde{\pi}_{X,Y}(\lambda_1, \frac{s_1-\lambda_1}{b} + s_2) - \frac{1}{(s_1+s_2+\frac{\mu+\lambda_1^A}{b})} \\ & \times \tilde{\pi}_{X,Y}(\lambda_1, s_1+s_2-\lambda_1)\Big].\end{aligned} \tag{5.38}$$

In order to obtain the explicit solutions, we notice that extra information is required. The following theorem specifies several boundary values to provide that information.

Theorem 5.5.1 *Let $\pi(0,0)$, $\pi(x,x)$ and $\pi(x,y)$ denote the limiting distributions of the states of an APQ under $M/M/1$ discipline. Further, let λ_i, $i = 1, 2$ be the Poisson arrival rates of class-i customers and μ be their common Exponential service rate. Without loss of generality, we assume the accumulating priority rates for class-1 and class-2 customers are set to 1 and b.*
The subsequent three rules will apply.

1. $\pi(0,0) = 1-\rho$
2. $\tilde{\pi}_X(0) = \rho_2 + \rho_1 b$
3. $\tilde{\pi}_{X,Y}(0,0) = \rho_1(1-b)$

Proof Lemma 4.2 in [10] states that a class-1 customer becomes accredited at rate $\lambda_1(1-b)$ when the queue is not empty. Class-1 customers arrive at rate λ_1, therefore the probability that they become accredited is $(1-b)$, while they arrive during a busy period. As a result the probability that they enter service as an un-accredited customer is b.

We also know that both classes of customers are arriving according to a Poisson process and therefore see time averages upon arrival. Therefore the stationary probability that a class-1 customer arrives to a busy queue and and begins service as an unaccredited customer is $\rho_1 b$. The probability that this customer starts service as an accredited customer is $\rho_1(1-b)$. Therefore the probability that a customer arrives to a busy queue and starts the service as an unaccredited customer is $\rho_2 + \rho_1 b$, since all class-2 customers start service as unaccredited.

We start by calculating $\tilde{\pi}_X(0)$:

$$\tilde{\pi}_X(0) = -(1-\rho) + \frac{(\frac{\mu+\lambda}{b})\tilde{\pi}_{X,Y}(\lambda_1, -\lambda_1)}{\rho\frac{\mu+\lambda_1^A}{b} - (\lambda_1 + \frac{\lambda_2}{b})}.$$

Consequently,

$$\begin{aligned}\tilde{\pi}_{X,Y}(\lambda_1, -\lambda_1) &= \frac{(\lambda_1 b - \lambda_1 + \mu)(\lambda_1(1-b)(1+\rho))}{\mu(\lambda+\mu)} \\ &= \rho_1(1-b)(1-\rho_1(1-b)).\end{aligned} \tag{5.39}$$

If we calculate $\tilde{\pi}_{X,Y}(0,0)$ we will have:

$$\begin{aligned}\tilde{\pi}_{X,Y}(0,0) = (1 - \frac{\mu}{\mu+\lambda_1^A})(-\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b})) + \tilde{\pi}_{X,Y}(\lambda_1, -\frac{\lambda_1}{b}) \\ - \frac{\mu}{\mu+\lambda_1^A}\tilde{\pi}_{X,Y}(\lambda_1, -\lambda_1).\end{aligned}$$

As a result,

$$\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\lambda_1}{b}) = \frac{\lambda_1^A}{\mu+\lambda_1^A}(\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) + 2). \tag{5.40}$$

If we assume that $\tilde{\pi}_{X,Y}(s_1, .)$ is a smooth function in λ_1, i.e. it is differentiable and both right and left limits converge to the actual value of function at λ_1, we will be able to find the limits and the value of function at critical points. Now let

$$\begin{aligned}\tilde{\pi}_{X,Y}(\lambda_1, s_2) &= \lim_{s_1 \to \lambda_1} \tilde{\pi}_{X,Y}(s_1, s_2) \\ &= \frac{-\mu\lambda_1(1-b)}{b^2(s_2 + \frac{\mu+\lambda_1}{b})^2}\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) + \frac{\mu\lambda_1}{b} \\ &\times \Big[\frac{d}{ds_1}\frac{\tilde{\pi}_{X,Y}(\lambda_1, \frac{s_1-\lambda_1}{b} + s_2)}{s_2 + \frac{\mu+s_1}{b}} - \frac{d}{ds_1}\frac{\tilde{\pi}_{X,Y}(\lambda_1, s_1+s_2-\lambda_1)}{s_1 + s_2 + \frac{\mu+\lambda_1^A}{b}}\Big]_{s_1=\lambda_1} \\ &= \frac{-\mu\lambda_1(1-b)}{b^2(s_2 + \frac{\mu+\lambda_1}{b})^2}\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) + \frac{\mu\lambda_1}{b} \\ &\times \Big[\frac{\frac{d}{ds_2}\tilde{\pi}_{X,Y}(\lambda_1, s_2)(s_2 + \frac{\mu+\lambda_1}{b})(1/b-1) - \tilde{\pi}_{X,Y}(\lambda_1, s_2)(1/b-1)}{(s_2 + \frac{\mu+\lambda_1}{b})^2}\Big].\end{aligned} \tag{5.41}$$

The foregoing equation can be rearranged as

$$((bs_2 + \mu + \lambda_1)^2 - (b-1)\mu\lambda_1)\tilde{\pi}_{X,Y}(\lambda_1, s_2) = \mu\lambda_1(b-1)(\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) - \frac{d}{ds_2}\tilde{\pi}_{X,Y}(\lambda_1, s_2)(s_2 + \frac{\mu+\lambda_1}{b})). \tag{5.42}$$

The foregoing equation is a linear first order differential equation in the form of $y' + p(s_2)y = g(s_2)$, where $y = \tilde{\pi}_{X,Y}(\lambda_1, s_2)$ and,

$$p(s_2) = \frac{b((bs_2 + \mu + \lambda_1)^2 - (b-1)\mu\lambda_1)}{((b-1)\mu\lambda_1)(bs_2 + \mu + \lambda_1)},$$

$$g(s_2) = b\frac{\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b})}{bs_2 + \mu + \lambda_1}.$$

By solving this equation, we can have a solution for $\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b})$, $\tilde{\pi}_{X,Y}(\lambda_1, s_1 + s_2 - \lambda_1)$ and $\tilde{\pi}_{X,Y}(\lambda_1, \frac{s_1-\lambda_1}{b} + s_2)$; which, finally give solutions for the Laplace transforms of the limiting distributions. In order to solve the differential equation, we define $\mu(x) = \exp[\int^x p(y)dy]$;

$$\begin{aligned}
\mu(x) &= \exp[\int^x \frac{b((bx + \mu + \lambda_1)^2 - (b-1)\mu\lambda_1)}{((b-1)\mu\lambda_1)(by + \mu + \lambda_1)}dy] \\
&= \exp[\frac{b}{(b-1)\mu\lambda_1}\int^x by + \mu + \lambda_1 dy - \int^x \frac{b}{by + \mu + \lambda_1}dy] \\
&= \exp[\frac{b}{(b-1)\mu\lambda_1}(\frac{bx^2}{2} + (\mu + \lambda_1)x) - ln(bx + \mu + \lambda_1)] \\
&= \frac{e^{\frac{b}{(b-1)\mu\lambda_1}(\frac{bx^2}{2} + (\mu+\lambda_1)x)}}{bx + \mu + \lambda_1}.
\end{aligned} \tag{5.43}$$

Therefore,

$$\begin{aligned}
\tilde{\pi}_{X,Y}(\lambda_1, x) &= y(x) \\
&= \frac{1}{\mu(x)}[\int^x \mu(y)b\frac{\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b})}{by + \mu + \lambda_1}dy + C].
\end{aligned} \tag{5.44}$$

Let $D = \int^x \mu(y)b\frac{\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b})}{by+\mu+\lambda_1}dy$, then we have:

$$
\begin{aligned}
D &= b\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) \int^x \frac{e^{\frac{b}{(b-1)\mu\lambda_1}(\frac{by^2}{2}+(\mu+\lambda_1)y)}}{(by+\mu+\lambda_1)^2} dy \\
&= b\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) \int^x \frac{e^{\frac{(by+(\mu+\lambda_1))^2-(\mu+\lambda_1)^2}{2(b-1)\mu\lambda_1}}}{(by+\mu+\lambda_1)^2} dy \\
&= b\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) e^{-\frac{(\mu+\lambda_1)^2}{2(b-1)\mu\lambda_1}} \int^x \frac{e^{\frac{(by+(\mu+\lambda_1))^2}{2(b-1)\mu\lambda_1}}}{(by+\mu+\lambda_1)^2} dy. \qquad (5.45)
\end{aligned}
$$

The last integral in (5.45) is $\frac{1}{b}\int^x \frac{e^{\frac{u^2}{c}}}{u^2} du$ if we let $u = by+(\mu+\lambda_1)$ and $c = 2(b-1)\mu\lambda_1$. Therefore, we can continue (5.44) as,

$$
\tilde{\pi}_{X,Y}(\lambda_1, x) = \frac{1}{\mu(x)}[\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) e^{-\frac{(\mu+\lambda_1)^2}{2(b-1)\mu\lambda_1}} \int^x \frac{e^{\frac{u^2}{c}}}{u^2} du + C]. \qquad (5.46)
$$

Hence, by employing the two equations (5.39) and (5.40) we will be able to solve for the unknown values of C and $\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b})$ in (5.46).

If we let $K(x) = \int^x \frac{e^{\frac{u^2}{c}}}{u^2} du$, then from (5.39) we have

$$
\begin{aligned}
\rho_1(1-b)(1-\rho_1(1-b)) &= \tilde{\pi}_{X,Y}(\lambda_1, -\lambda_1) \\
&= \frac{1}{\mu(-\lambda_1)}[\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) e^{-\frac{(\mu+\lambda_1)^2}{2(b-1)\mu\lambda_1}} K(-\lambda_1) + C]. \qquad (5.47)
\end{aligned}
$$

From (5.40) we get,

$$
\begin{aligned}
&\frac{\lambda_1^A}{\mu+\lambda_1^A}(\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) + 2) = \tilde{\pi}_{X,Y}(\lambda_1, -\frac{\lambda_1}{b}) \\
&= \frac{1}{\mu(-\frac{\lambda_1}{b})}[\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) e^{-\frac{(\mu+\lambda_1)^2}{2(b-1)\mu\lambda_1}} K(-\frac{\lambda_1}{b}) + C]. \qquad (5.48)
\end{aligned}
$$

After solving these equations we will have:

$$
C = [\frac{\rho_1^A(1-(\rho_1^A)^2)B\mu(-\lambda_1) + 2\rho_1^A}{(1+\rho_1^A)B\mu(-\lambda_1) + 1}]\mu(-\frac{\lambda_1}{b}), \qquad (5.49)
$$

where $B = \frac{\rho_1^A e^{\frac{(\mu+\lambda_1)^2}{2(b-1)\mu\lambda_1}}}{(1+\rho_1^A)K(-\frac{\lambda_1}{b})} - [\mu(\frac{-\lambda_1}{b})]^{-1}$ and, finally

$$\tilde{\pi}_{X,Y}(\lambda_1, -\frac{\mu+\lambda_1}{b}) = \frac{C(1+\rho_1^A) - 2\rho_1^A\mu(-\frac{\lambda_1}{b})}{\rho_1^A\mu(-\frac{\lambda_1}{b}) - (1+\rho_1^A)e^{-\frac{(\mu+\lambda_1)^2}{2(b-1)\mu\lambda_1}}K(-\frac{\lambda_1}{b})}. \tag{5.50}$$

At this point, we are able to define $\tilde{\pi}_{X,Y}(\lambda_1, x)$ as in (5.46). This will give rise to explicit solutions for Equation (5.36) and Equation (5.38). Thus we have the explicit form for our LST functions $\tilde{\pi}_{X,Y}(s_1, s_2)$ and $\tilde{\pi}_X(s)$.

At this stage, we have determined the LST of the stationary accumulated priorities at the time points that customers move into service. The numerical Gaver-Stehfest Algorithm can be used to numerically invert the Laplace transforms and obtain the actual steady state distribution functions in a classical APQ setting when required.

In the following section we will discuss the connection between LST of the stationary accumulated priorities and the stationary waiting time distributions.

5.6 Waiting time distributions

We return now to the discussion of the waiting time distributions. In this section, we establish the necessary relations to link the results we acheived in the previous section to the LST of the waiting time distributions in a classical APQ. Once we have the LST for the stationary accumulated priorities at the time points that customers move into the service, we are also immediately able to derive the LST for the stationary waiting times by a re-scaling of the arguments. The underlying reason is that when a customer of class-i with v accumulated priority enters into service, the waiting time required for this person to gain v credit had been v/b_i as also discussed in Stanford et al. (2014).

Stanford et al. (2014) (Equation [30], page 315) have derived the LST of the stationary accumulated priority of the non-accredited customers at the time that they enter service, conditional on its being positive, as $\tilde{V}^{(2)}(s)$. They have employed that result to obtain the stationary waiting time distribution for class-2 customers as the weighted sum of the LSTs of zero and $\tilde{V}^{(2)}(\frac{s}{b_2})$ (Equation [31], page 315). Applying the same argument, in the following theorem, we derive the LST for the waiting time distribution of class-2 customers.

$$\tilde{W}^{(2)}(s) = (1-\rho) + \rho\tilde{V}^{(2)}(s/b_2). \tag{5.51}$$

Theorem 5.6.1 *Let $\tilde{V}^{(2)}(s)$ be the LST of the stationary accumulated priority of the non-accredited customers at the time that they enter service, conditional on its being positive, in a two class APQ under $M/M/1$ discipline with parameters $b_1 = 1$, $b_2 = b$ and λ_1. Also let $\tilde{\pi}_X(s)$ be the LST*

of the stationary accumulated priority of the non-accredited customers at the time of entrance int service. We will have the two following relations:

1.
$$\tilde{V}^{(2)}(s) = \frac{\tilde{\pi}_X(s)}{\rho_2 + \rho_1 b} \tag{5.52}$$

2.
$$\tilde{W}^{(2)}(s) = (1-\rho) + \rho \frac{\tilde{\pi}_X(s/b_2)}{\rho_2 + \rho_1 b}, \tag{5.53}$$

where $\rho_1 = \lambda_1/\mu$ and $\rho_2 = \lambda_2/\mu$.

Proof The difference between $\tilde{V}^{(2)}(s)$ and $\tilde{\pi}_X(s)$ is that the former is the conditional LST of the stationary accumulated priorities of the non-accredited customers and the latter is the unconditional (joint) distribution of the non-accredited class-1 and class-2 customers.
A class-1 customer becomes accredited at rate $\lambda_1(1-b)$ (Lemma 4.2 in Stanford et al. (2014)). Therefore, the probability that an individual class-1 customer, arriving during a busy period, enters service while unaccredited is b. As a result, the probability that a non-accredited customer enters service with positive credit (i.e. after waiting for positive units of time)is $\rho_2 + \rho_1 b$. Thus, based on the conditional probability rule we can write $\tilde{V}^{(2)}(s) = \frac{\tilde{\pi}_X(s)}{\rho_2+\rho_1 b}$.

Since a class-2 customer either finds the server empty at the arrival moment or waits in the system for time v/b_2, the LST of the stationary waiting time for class-2 customers can be obtained by the weighted sum of the LSTs of zero and $\tilde{\pi}_X(s/b)$. Therefore by substituting Equation (5.52) in (5.51) we will obtain (5.53).

Applying similar argument we are ale to derive the waiting time distributions for class-1 customers according to the following theorem,

Theorem 5.6.2 *Let the LST of the distribution of the priority of a class-*1 *customer when it enters service, conditional on this being positive, be* $\tilde{V}^{(2)}(s)$ *(Equation 36 on page 316 in Stanford et al. (2014)). Also, let* $\tilde{\pi}_{X,Y}(s_1, s_2)$ *be the LST of the joint distribution of the* $M_1(t)$ *and* $M_2(t)$ *when the customer enters service (derived in the previous section). The LST of the stationary waiting time for class-*1 *customer is as follows:*

1.
$$\tilde{V}^{(1)}(s) = b\frac{\tilde{\pi}_X(s)}{\rho_2 + \rho_1 b} + (1-b)\frac{\tilde{\pi}_{X,Y}(s,0)}{\rho_1(1-b)} \tag{5.54}$$

2.
$$\tilde{W}^{(1)}(s) = (1-\rho) + \rho[b\frac{\tilde{\pi}_X(s/b_1)}{\rho_2 + \rho_1 b} + (1-b)\frac{\tilde{\pi}_{X,Y}(s/b_1,0)}{\rho_1(1-b)}]. \tag{5.55}$$

Proof A class-1 customer arrives to an empty queue with probability $(1-\rho)$. If it arrives to a non-empty queue, it enters service non-accredited with probability $b_2/b_1 = b$; and, the LST of its stationary accumulated priority on entering service is according to (5.52). On the other hand, if a class-1 customer arrives to a non-empty queue, it enters service as an accredited customer with probability $1-b$ and its stationary accumulated priority LST on entering service would be $\frac{\tilde{\pi}_{(X,Y)}(s,0)}{\rho_1(1-b)}$.

Therefore, the LST of the distribution of the priority of this customer when it enters service, conditional on being positive is according to (5.54).

The LST of the waiting time is obtained by substituting (5.54) in $\tilde{W}^{(1)}(s) = (1-\rho)+\rho\tilde{V}^{(1)}(s/b_1)$ (Equation[37] on page 316 Stanford et al. (2015)).

At this stage we are able to find the LST of the waiting time for class-1 and class-2 customers in an APQ under $M/M/1$ setting with one server.

5.7 Conclusions and future work

In the third chapter of this thesis we presented an algorithm to derive the waiting time distribution for the lowest priority class in an Affine APQ under some specific assumptions. Unlike APQs, in an affine APQ a positive credit is assigned to each class of customers upon their arrivals. Without loss of generality in a two-class setting, we assume the initial credit for the lower priority class is 0 and for the higher priority class is the positive value a. Aiming to derive the wait time distributions for both classes of customers in an Affine APQ setting under M/M1 discipline we analysed the corresponding bi-variate Maximum Priority Process in this setting, as having the exact values of this process at the specific instant of a new service commencements leads to the derivation of the waiting times for each customer.

Therefore, we derived the explicit solution to the LST of the stationary accumulated priority at the time of entrance when $a = 0$ (Classical APQ setting). Finally, we were able to employ these results to derive the LST of the waiting time distributions in an APQ with 2 classes of customers under $M/M/1$ discipline. This new approach could be used as a tool to solve more general cases.

Since in the present work, the analysis of the maximum priority process is done for the Affine APQs, it could be used for studying the LST of the waiting times in an Affine APQ as future research. We have derived the explicit solution for the LST of the steady states when $a = 0$; an extension to this work when $a > 0$ could be a nice problem to solve as a future research.

Finally this work presents a new approach to determine the LST of the waiting time distribution for higher and lower class customers in an APQ with 2-classes of customers under the $M/M/1$ discipline.

Chapter 6

Conclusion and future work

6.1 Main contributions

1. The third chapter "Discrete time Markov chain algorithm for short time predictions in an Emergency Department" pertains to the near future predictions in an Emergency Department (ED) to assist the decision makers with planning the best interventions for their system. Short-run predictions of ED censuses are particularly important for efficient allocation and management of ED resources. Initially, we investigated both regression and time-series based forecasting methods to identify an appropriate forecasting model to accurately predict ED arrivals and discharges in short term. Considering the variation in arrival pattern and service requirements, we applied and compared three models which best described our data.

In our study, we modeled ED changes based on a Discrete Time Markov Chain (DTMC) algorithm we introduced. We presented estimations for short term (hourly) ED censuses at each time point and provided hourly predictions up to 24 hours in a day which can potentially provide suggestions to ED managers by constructing numerical analysis on how to prevent over-crowding in their system.

We illustrated our approach using 22 months of data obtained from the ED of a large academic medical center in Ontario. Our three models were validated and compared in accuracy and functionality based on MSE and correlation.

2. The fourth chapter "The Lowest Priority Waiting Time Distribution in the Affine and the Delayed Accumulating Priority Queues" introduced Affine and Delayed variants of the APQ inspired by health care applications. In the Affine APQ setting the high-acuity patients receive positive a credits upon arrival and accumulate credits so long as they wait in the queue with a higher accumulation rate as compared to the lower-acuity patients. In the delayed APQ setting, the low-acuity patients do not accumulate priority over time until the period of delay has been reached. Even though the motivations for the affine and the delayed variants appear to describe differing priority accumulation mechanisms, we established that they are, in fact, equivalent. It was established that the waiting time distribution of the lower-priority class of customers in both variants is identical to those of the lower class in a classical priority queue, up to the time

threshold. Beyond that time point, the waiting time behaviour resembles that of a non-affine APQ. We were able to exploit these facts to come up with an algorithm for the determination of the waiting time distribution for those customers who experience waits in excess of the time threshold. Numerical examples were presented to illustrate the trends we observed.

The second contribution of this work is related to the trend for health care systems to respond to so-called "Key Performance Indicators (KPIs)". The KPI approach specifies, for each class of customers both a time target for customers to commence service, and a compliance probability indicating the proportion of customers that meet the target. The main problem with working solely on the basis of KPIs is that no consequence is specified for customers who miss their target, when in fact a customer who misses their KPI target maybe is of greater if not lesser importance. Therefore, we investigated the question of how to determine the optimal accumulation rate for lower-acuity patients as a function of queue's occupancy level at different delay time period in the delayed variant of APQ (equivalently, at different a levels for the Affine APQ). We addressed this question by solving for the optimum accumulation rate, b, so that a decision maker can select any priority accumulation rate they like, so long as low-acuity patients can meet their corresponding KPIs. We illustrated the results through some numerical examples.

3. The fifth chapter "The bivariate Maximum Priority Process in an Affine APQ" introduced the bivariate Maximum Priority process in an Affine APQ setting in a two-class queue. We noticed that this process, at the times at which customers move into service represents the exact credit value of the customer who starts the service. This property can ultimately be related to the waiting time distributions for both acuity classes. We observed that this process at the instant that a new service commences is a Markov process for which we identified five possible state sets and derived the LST of stationary distributions.

Due to difficulties in solving the general recursive LST functions in the Affine APQ setting, we we let the affine parameter $a = 0$, and derived the explicit solutions for the LSTs. Then, we managed to link the results with the LST of the stationary accumulated priorities at the time points that customers move into service. Therefore, we also obtained the LST for the stationary waiting times by investigating the appropriate re-scaling of the arguments. Therefore, this study introduces a new approach to study classical APQ's with a strong potential to be expanded to the Affine APQ.

6.2 Future work

Some extensions to our work are possible. We highlight below some of them which can be considered as future research.

1. In chapter three we proposed an algorithm to predict hourly census in an ED. One obvious extension with the available data could be investigating the hourly fluctuations of the system under different initial-work-load scenarios to identify a rule to the optimum intervention moment and best action plan to avoid overcrowding.

2. In case of having access to a data set with more detailed information on the number of staff, bed availability or other resources at each time point, a Markov decision process can be used to dynamically make the optimum decision (chose the best action/intervention) to ultimately reduce the congestion in the system. Actions could include managing the number of nurse staffing, doctor staffing or temporarily allocation of other available resources.

3. In chapter five, we analysed the Maximum priority process in the Affine APQ but solved for the waiting time LSTs in a specific case of $a = 0$ (i.e. classical APQ). As mentioned in the discussion, the long run probability of $M_2(t) < a$ would be the main key to solve this problem for the general Affine case. Identifying this key probability will be useful for the ultimate resolution of the bivariate process when $a > 0$. In order to obtain the waiting time distribution for the higher priority class, one needs to distinguish between periods when $M_2(t) \leq a$ from $M_2(t) > a$, as this affects who a tagged arrival from the higher class would wait for.

Bibliography

[1] Boyle A., Beniuk K., Higginson I., and Atkinson P. Emergency department crowding: Time for interventions and policy evaluations. *Emerg. Med. Int.*, 2012.

[2] Carpinone A., Giorgio M., Langella R., and Testa A. Markov chain modeling for very-short-term wind power forecasting. *Electric Power Systems Research*, 122:152–158, 2015.

[3] Cobham A. Priority assignment in waiting line problems. *Journal of the Operations Research Society of America*, 2:70–76, 1954.

[4] Fajardo V. A. *A Generalization of $M/G/1$ priority models via accumulating priority*. PhD thesis, The University of Waterloo, 2015.

[5] Fajardo V. A. and Drekic S. Waiting time distributions in the preemptive accumulating priority queue. *Methodology and Computing in Applied Probability*, 2017.

[6] Schuster A. II. On the periodicities of sunspots. *Philosophical Transactions of the Royal Society*, 1906.

[7] Shamshad A., Bawadi M. A., Hussin W., Majid T. A., and Sanusi S. A. M. First and second order Markov chain models for synthetic generation of wind speed time series. *Energy*, 30, 2005.

[8] Sharif B. A., Stanford A. D., Taylor P., and Ziedins I. A multi-class multi-server accumulating priority queue with application to healthcare. *Operations research for Health Care*, 2014.

[9] Stanford D. A., Pagurek B., and Woodside C. M. Optimal prediction of times and queue length in the GI/M/1 queue. *INFORMS*, 31:322–337, 1983.

[10] Stanford D. A., Taylor P., and Ziedins I. Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 2014.

[11] Wait Time Alliance. Report card on wait times in Canada. From the website. http://www.waittimealliance.ca/wta-reports/2014.

[12] Sharif A. B. *Probability models for healthcare operations with application to emergency medicine*. PhD thesis, Western University, 2016.

[13] Raible C. C., Bischof G., Fraedrich K., and Kirk E. Statistical single-station short-term forecasting of temperature and probability of precipitation: Area interpolation and NWP combination. *Weather and Forecasting*, 1999.

[14] Gross D. and Harris C. M. *Fundamentals of queueing theory*. John Wiley & Sons, 1974.

[15] Boyce W. E. and DiPrima R. C. *Elementary differential equations and boundary value problems*. Wiley, New York, 1965.

[16] Crabtree B. F., Ray S. C., Schmidt P. M., O'Connor P. J., and Schmidt D. D. The individual over time: time series applications in health care research. *Journal of Clinical Epidemiology*, 43 3:241–60, 1990.

[17] Dreyer J. F, McLeod S. L., Anderson C. K., Carter M. W., and Zaric G. S. Physician workload and the Canadian Emergency Department Triage and Acuity Scale: the Predictors of Workload in the Emergency Room (POWER) Study. *Canadian Journal of Emergency Medicine*, 11:321–329, 2009.

[18] Yu G., Hu J., Zhang C., Zhuang L., and Song J. Short-term traffic flow forecasting based on Markov chain model. *Proc. IEEE Intell. Vehicles Symp.*, pages 208–212, 2003.

[19] Kesten H. and Runnenberg J. Th. Priority in waiting line problems. *Nederlandse Akademie van Wetenschappen. Proceedings. Series A. Indagationes Mathematicae.*, 1957.

[20] Stehfest H. Algorithm 368: Numerical Inversion of Laplace Transforms. *Communications of the ACM*, 1970.

[21] Adan I. and Haviv M. Conditional ages and residual service times in the M/G/1 queue. *Stochastic Models*, 2009.

[22] Bullard M. J., Unger B., Spence J., Grafstein E., and CTAS National Working Group. Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) adult guidelines. *CJEM*, 10:136–151, 2008.

[23] Côté M. J., Smith M. A., Eitel D. R., and Akçali E. Forcasting emergency department arrivals: A tutorial for emergency department directors. *Hospital Topics*, 91(1):9–19, 2013.

[24] Flottemesch T. J., Gordon B. D., and Jones S. S. Developing a formal model of emergency department census and defining operational efficiency. *Acad Emerg Med*, 2007.

[25] Hsu J. A continuation of delay-dependent queue disciplines. *Operations Research*, 1970.

[26] Fraedrich K. and Muller K. On single station forecasting: Sunshine and rainfall Markov chains. *Contrib. Atmos. Phys.*, 56:108–134, 1983.

[27] Fraedrich K. and Leslie L. M. Combining predictive schemes in short-term forecasting. *Monthly weather review*, 1987.

[28] Kleinrock L. A delay dependent queue discipline. *Naval Research Logistic Quarterly*, 11:329–341, 1964.

[29] Kleinrock L. *Queueing Systems*, volume 1: Theory. Wiley, New York, 1975.

[30] Kleinrock L. and Finkelstein R. Time dependent priority queues. *Operations Research*, 15:104–116, 1967.

[31] Haviv M. and Ravner L. Strategic bidding in an accumulating priority queue: equilibrium analysis. *Computer Science and Game Theory*, 2016.

[32] Ross S. M. *Stochastic Processes*, chapter 4. Wiley, 1995.

[33] Channouf N., L'Ecuyer P., Ingolfsson A., and Avramidis A. N. The application of forcasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Manage Sci*, 10:25–45, 2007.

[34] Li N. *Recent advances in Accumulating Priority Queues*. PhD thesis, Western University, 2015.

[35] Li N. and Stanford D. Multi-server accumulating priority queues with heterogeneous servers. *European Journal of Operational Research*, 252(3):866–878, 2016.

[36] Li N., Stanford A. D., Taylor P., and Ziedins I. Nonlinear accumulating priority queues with equivalent linear proxies. *Operations Research*, 2017.

[37] Soyiri I. N. and Reidpath D. D. An overview of health forecasting. *Operations research for Health Care*, 2012.

[38] Kella O. and Ravner L. Lowest priority waiting time distribution in an accumulating priority lvy queue. *Operations Research Letters*, 45(1):40–45, 2017.

[39] Miró O., Antonio M. T., Jiménez S., De Dios A., Sánchez M., Borrás A., and Millá J. Decreased health care quality associated with emergency department overcrowding. *Eur J Emerg Med.*, 6(2):105–107, 1999.

[40] Gaver D. P. Observing stochastic processes,and approximate transform inversion. *Operations Research*, 1966.

[41] Milner P. Forecasting the demand on accident and emergency departments in health districts in the Trent region. *Statistics in medicine*, 7(10):0611072, 1988.

[42] Sethi S. P. and Stroger G. A theory of rolling horizon decision making. *Annals of Operations Research*, 29(1), 1991.

[43] Sethi S. P., Yan H., and Zhang H. Inventory and supply chain management with forecast updates. *International series in operations research & management science*, 81, 2005.

[44] Li QL. *Markov Chains on Continuous State Space. In: Constructive Computation in Stochastic Models with Applications.*, chapter 5. Springer, Berlin, Heidelberg, 2010.

[45] Aspline B. R., Flottemesch T. J., and Gordon B. R. Developing models for patient flow and daily surge capacity research. *Academic emergency Medicine*, 2006.

[46] Broyel J. R., Cochran J. K., and Montgomery D. C. A statistical Marcov chain approximation of transient hospital inpatient inventory. *European Journal of Operations Research*, pages 1645–1654, 2010.

[47] Broyel J. R., Cochran J. K., and Montgomery D. C. A Markov decision process to dynamically match hospital inpatient staffing to demand. *IIE Transactions on Healthcare systems Engineering*, pages 116–130, 2011.

[48] Eitel D. R., Rudkin S. E., Malvehi L. A., Killeen J. P., Pines J. M., Gadd C. S., and Aronsky D. Imroving service quality by understanding emergency department flow: A white paper and position statement paper for the American academy of emergency medicine. *Journal of Emergency medicine*, 39(1):70–79, 2010.

[49] Gabriel K. R. and Neumann J. A markov chain model for daily rainfall occurrence. *Quartely Journal of the Royal Meteorological Society*, 1962.

[50] Hoot N. R. and Aronsky D. An early warning system for overcrowding in the emergency department. *Proc AMIA Annu Fall Symp*, pages 339–343, 2006.

[51] Hoot N. R. and Aronsky D. Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions. *Annals of Emergency Medicine*, 52(2):126–136, 2008.

[52] Hoot N. R., LeBlanc L. J., Jones I., Levin S. R., Zhou C., Gadd C. S., and Aronsky D. Forecasting emergency department crowding: a discrete event simulation. *Annals of Emergency medicine*, 52(2):116–125, 2008.

[53] Jackson J. R. Some problems in queueing with dynamic priorities. *Naval Research Logistics Quarterly*, 1960.

[54] Jackson J. R. Queues with dynamic priority discipline. *Management Science*, 1961.

[55] Jackson J. R. Waiting-time distributions for queues with dynamic priorities. *Naval Research Logistics Quarterly*, 1962.

[56] Jones S. S., Thomas A., Evans R. S., Welch S. J., Haug P. J., and Snow G. L. Forecasting daily patient volumes in the emergency department. *Acad Emerg Med*, 15(2):159–170, 2008.

[57] Jones S. S., Evans R. S., Allen T. L., Thomas A., Haug P. J., welch S. J., and Snow G. L. A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of Biomedical Informatics*, 42:123–139, 2009.

[58] Oredsson S., Jonsson H., Rognes J., Lind L., Göransson K. E., Ehrenberg A., Asplund K., Castrén M., and Farrohknia N. A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scand J Trauma Resusc Emerg Med*, 19:19–43, 2011.

[59] Robert H. Shumway and David S. Stoffer. *Time series analysis and its applications: with R examples*, chapter 5. Springer, 2011.

[60] Yechiali U. and Kella O. Waiting times in the non-preemptive priority $M/M/c$ queue. *Communications in Statistics. Part C: Stochastic Models*, 1(2):257–262, 1985.

[61] Conway R. W., Maxwell W. L., and Miller L. W. *Theory of scheduling*. Addison-Wesley., 1967.

[62] Sun Y., Heng B. H., Seow Y. T., and Seow E. Forcasting daily attendances at an emergancy department to aid resource planning. *BMC Emergency medicine*, 9, 2009.

[63] Krougly Z., Davison M., and S Aiyar. The role of high precision arithmetic in calculating numerical Laplace and inverse Laplace transforms. *Applied Mathematics*, 2017.

Appendix A

Additional materials in Chapter 5

The process of deriving the LST for all of the limiting distributions in Chapter 5, is very tedious. In this process after identifying possible state sets, we derived the corresponding kernel densities. A lot of mathematical details were involved in the solution of the integral equations. However, for the sake of readability, I provide further details of derivations not in the main text but in this part. Some necessary derivations which have appeared in final results, will be presented in more detail as follows.

A.1 Checking for the pdf assumption in kernel densities in Section 5.3.2

In order to verify whether the transition probabilities from state (sets) form a density function, we compute the total probability value on the whole admissible region. Therefore, we show that the total probability adds up to one.

I. state $(a, 0)$:

$$
\begin{aligned}
& P_{(a,0)\to(a,0)} + \int_{y=0}^{a} f_{\{(a,0)\to(a,y),y\le a\}} dy + \int_{x=0}^{\infty} f_{\{(a,0)\to(x,x),x\ge a\}} dx + \int_{x=a}^{\infty}\int_{y=b(x-a)}^{x} \\
& \quad f_{\{(a,0)\to(x,y)\}} dy dx \\
&= \frac{\mu}{\mu+\lambda}\Big(1 + \frac{\lambda_2}{\mu+\lambda_1}\big(1 - e^{-\frac{a}{b}(\mu+\lambda_1)}\big) + \frac{\lambda_1 b + \lambda_2}{\mu+\lambda_1^A} e^{-\frac{a}{b}(\mu+\lambda_1)}\Big) + \frac{\lambda_1\mu}{\mu+\lambda_1}\Big(\frac{1}{\mu} - \frac{b}{(\mu+\lambda_1^A)} e^{-\frac{a}{b}(\mu+\lambda_1)}\Big) \\
&= \frac{\mu}{\mu+\lambda}\Big(1 + \frac{\lambda_2}{\mu+\lambda_1}\Big) + \frac{\lambda_1}{\mu+\lambda_1} + \Big(\frac{\mu}{\mu+\lambda}\Big(\frac{\lambda_1 b + \lambda_2}{\mu+\lambda_1^A} - \frac{\lambda_2}{\mu+\lambda_1}\Big) - \frac{\mu\lambda_1 b}{(\mu+\lambda_1)(\mu+\lambda_1^A)}\Big) e^{-\frac{a}{b}(\mu+\lambda_1)} \\
&= 1. \qquad \text{(A.1)}
\end{aligned}
$$

II. state (a, y); $y < a$:

$$
\begin{aligned}
&P_{(a,w)\to(a,0)} + \int_{y=0}^{w} f_{\{(a,w)\to(a,y),y\le w\}}dy + \int_{y=w}^{a} f_{\{(a,w)\to(a,y),w\le y\}}dy + \int_{x=a}^{\infty} f_{\{(a,w)\to(x,x),x\ge a\}}dx \\
&\quad + \int_{x=a}^{\infty}\int_{y=b(x-a)+w}^{x} f_{\{(a,0)\to(x,y)\}}dydx \\
&= \frac{\mu}{\mu+\lambda}(1 + \frac{\lambda_2}{\mu+\lambda_1}(1 - e^{-\frac{a-w}{b}(\mu+\lambda_1)}) + \frac{\lambda_1 b+\lambda_2}{\mu+\lambda_1^A}e^{-\frac{a-w}{b}(\mu+\lambda_1)}) + \frac{\lambda_1\mu}{\mu+\lambda_1}(\frac{1}{\mu} - \frac{b}{(\mu+\lambda_1^A)}e^{-\frac{a-w}{b}(\mu+\lambda_1)}) \\
&= \frac{\mu}{\mu+\lambda}(1 + \frac{\lambda_2}{\mu+\lambda_1}) + \frac{\lambda_1}{\mu+\lambda_1} + (\frac{\mu}{\mu+\lambda}(\frac{\lambda_1 b+\lambda_2}{\mu+\lambda_1^A} - \frac{\lambda_2}{\mu+\lambda_1}) - \frac{\mu\lambda_1 b}{(\mu+\lambda_1)(\mu+\lambda_1^A)})e^{-\frac{a-w}{b}(\mu+\lambda_1)} \\
&= 1. \qquad \text{(A.2)}
\end{aligned}
$$

III. state (x, x):

$$
\begin{aligned}
&P_{(\nu,\nu)\to(a,0)} + \int_{y=0}^{a} f_{\{(\nu,\nu)\to(a,y),y\le a\}}dy + \int_{y=a}^{\nu} f_{\{(\nu,\nu)\to(x,x),a\le x<\nu\}}dy \\
&\quad + \int_{x=\nu}^{\infty} f_{\{(\nu,\nu)\to(x,x),x\ge\nu\}}dx + \int_{x=\nu}^{\infty}\int_{y=b(x-\nu)+\nu}^{x} f_{\{(\nu,\nu)\to(x,\nu)\}}dydx \\
&= \frac{\mu}{\mu+\lambda}(1 + \frac{\lambda_1 b+\lambda_2}{\mu+\lambda_1^A}) + \mu e^{-\lambda_1\nu}(\frac{e^{\lambda_1\nu}}{\mu} - \frac{e^{\lambda_1\nu}}{\mu+\lambda_1^A}) \\
&= 1. \qquad \text{(A.3)}
\end{aligned}
$$

IV. state (x, y):

In order to verify if the probabilities add up to one, we consider both cases $a < w$ (i.e. when the maximum priority for class 2 customers, $M_2(t)$, at the service initiation moment, is more than a) and $w < a$ (i.e. when $M_2(t)$ is less than a when an accredited customer starts service) separately as follows:

a) When $a < w$:

$$
\begin{aligned}
&P_{(\nu,w)\to(a,0)} + \int_{y=0}^{a} f_{\{(\nu,w)\to(a,y),y<a<w\}}dy + \int_{y=a}^{w} f_{\{(\nu,w)\to(x,x),a\le x<w\}}dy \\
&\quad + \int_{x=w}^{\infty} f_{\{(\nu,w)\to(x,x),x\ge w\}}dx + \int_{y=w}^{\infty}\int_{x=y}^{\frac{y-w}{b}+\nu} f_{\{(\nu,w)\to(x,y)\}}dxdy \\
&= \frac{\mu}{\mu+\lambda}(1 + \frac{\lambda_1 b+\lambda_2}{\mu+\lambda_1^A})e^{-\lambda_1(\nu-w)} + 1 - \frac{\mu}{\mu+\lambda_1^A}e^{-\lambda_1(\nu-w)} \\
&= 1. \qquad \text{(A.4)}
\end{aligned}
$$

b) Similarly, when $w < a$:

$$
\begin{aligned}
&P_{(\nu,w)\to(a,0)} + \int_{y=w}^{a} f_{\{(\nu,w)\to(a,y),w<y<a\}}dy + \int_{y=0}^{w} f_{\{(\nu,w)\to(a,y),y<w<a\}}dy \\
&\quad + \int_{x=a}^{\infty} f_{\{(\nu,w)\to(x,x),w<a<x\}}dx + \int_{y=w}^{a}\int_{x=a}^{\frac{y-w}{b}+\nu} f_{\{(\nu,w)\to(x,y)\}}dxdy \\
&\quad + \int_{y=a}^{\infty}\int_{x=y}^{\frac{y-w}{b}+\nu} f_{\{(\nu,w)\to(x,y)\}}dxdy \\
&= \frac{\mu}{\mu+\lambda}e^{-\lambda_1(\nu-a)}\Big[\frac{\lambda_2}{\mu+\lambda_1} - \frac{\lambda_2}{\mu+\lambda_1}e^{\frac{\mu+\lambda_1}{b}(w-a)} + 1 + \frac{b}{\mu+\lambda_1^A}e^{(w-a)(\frac{\mu+\lambda_1}{b})}(\lambda_1+\frac{\lambda_2}{b})\Big] \\
&\quad + 1 + \frac{\mu}{\mu+\lambda_1}e^{-\lambda_1(\nu-a)}[e^{\frac{\mu+\lambda_1}{b}(w-a)} - 1] - \frac{\mu}{\mu+\lambda_1^A}e^{\lambda_1(a-\nu)}.e^{(\frac{\mu+\lambda_1}{b})(w-a)} \\
&= 1. \qquad \text{(A.5)}
\end{aligned}
$$

A.2 More details on Equations in Section 5.4

In this part, we shed more light on the details of the derivations and Lemmas presented in section 5.4.

1. **Proof for Lemma 5.4.2**: *The LST of the stationary distribution for state set* $(a, y); y < a$ *is according to:*

$$
\begin{aligned}
\tilde{\pi}_Y(s) = B(s)^{-1}\Big[&\frac{1-e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s+\frac{\mu+\lambda_1}{b}}(1-\rho) + \frac{1-e^{-(s-\frac{\lambda_2}{b})a}}{s-\frac{\lambda_2}{b}}[\rho(1-\rho)] + \frac{e^{\lambda_1 a}}{s-\frac{\lambda_2}{b}} \\
&\times\Big(e^{-(s-\frac{\lambda_2}{b})a}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,\frac{\lambda_2}{b}) - \frac{\mu+\lambda}{b(s+\frac{\mu+\lambda_1}{b})}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s)\Big) \\
&- \frac{e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s+\frac{\mu+\lambda_1}{b}}e^{\lambda_1 a}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big], \qquad \text{(A.6)}
\end{aligned}
$$

when $B(s) = \frac{b(1+\rho)}{\lambda_2} - \frac{1-e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s+\frac{\mu+\lambda_1}{b}} + \frac{1-e^{-(s-\frac{\lambda_2}{b})a}}{s-\frac{\lambda_2}{b}}$.

Proof We evaluate the integrals in Equation (5.26) according to the corresponding transition densities we have developed in Subsection 5.3.2. Doing so we will have the following

$$\frac{\mu+\lambda}{\mu}.\frac{b}{\lambda_2}\,\pi(a,y) = \pi(a,0)e^{-\frac{\mu+\lambda_1}{b}y} + \int_0^y \pi(a,w)e^{-\frac{\mu+\lambda_1}{b}(y-w)}dw + \int_y^a \pi(a,w) \times e^{-\frac{\lambda_2}{b}(w-y)}dw + \int_a^\infty \pi(\nu,\nu)e^{-\lambda_1(\nu-a)-\frac{\lambda_2}{b}(\nu-y)}d\nu + \int_a^\infty \int_w^{\frac{w}{b}+a} \pi(\nu,w) \times e^{-\lambda_1(\nu-a)-\frac{\lambda_2}{b}(w-y)}d\nu dw + \int_0^y \int_a^{\frac{w}{b}+a} \pi(\nu,w)e^{-\lambda_1(\nu-a)-\frac{\mu+\lambda_1}{b}(y-w)}d\nu dw + \int_y^a \int_a^{\frac{w}{b}+a} \pi(\nu,w)e^{-\lambda_1(\nu-a)-\frac{\lambda_2}{b}(w-y)}d\nu dw. \tag{A.7}$$

After taking Laplace transform of both sides of the above equation, we simplify the integrations according to the following relations:

$$\int_0^a \pi(a,0)e^{-sy}e^{-\frac{\mu+\lambda_1}{b}y}dy = \frac{1-e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s+\frac{\mu+\lambda_1}{b}}\pi(a,0), \tag{A.8}$$

$$\int_0^a e^{-sy}\int_0^y \pi(a,w)e^{-\frac{\mu+\lambda_1}{b}(y-w)}dwdy = \frac{1-e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s+\frac{\mu+\lambda_1}{b}}\tilde{\pi}_y(s) \tag{A.9}$$

$$\int_0^a e^{-sy}\int_y^a \pi(a,w)e^{-\frac{\lambda_2}{b}(w-y)}dwdy = \frac{1-e^{-(s-\frac{\lambda_2}{b})a}}{s-\frac{\lambda_2}{b}}(\tilde{\pi}_y(\frac{\lambda_2}{b}) - \tilde{\pi}_y(s)), \tag{A.10}$$

$$\begin{aligned}\int_0^a e^{-sy}\int_a^\infty \pi(\nu,\nu)e^{-\lambda_1(\nu-a)-\frac{\lambda_2}{b}(\nu-y)}d\nu dy &= \int_0^a e^{-sy}e^{\lambda_1 a+\frac{\lambda_2}{b}y}\tilde{\pi}_x(\lambda_1+\frac{\lambda_2}{b})dy \\ &= e^{\lambda_1 a}(\frac{1-e^{-(s-\frac{\lambda_2}{b})a}}{s-\frac{\lambda_2}{b}})\tilde{\pi}_x(\lambda_1+\frac{\lambda_2}{b}),\end{aligned} \tag{A.11}$$

$$\begin{aligned}&\int_0^a e^{-sy}\int_0^y \int_a^{\frac{w}{b}+a} \pi(\nu,w)e^{-\lambda_1(\nu-a)-\frac{\mu+\lambda_1}{b}(y-w)}d\nu dwdy \\ &= e^{\lambda_1 a}\int_0^a e^{w(\frac{\mu+\lambda_1}{b})}\int_a^{\frac{w}{b}+a} e^{-\lambda_1\nu}\pi(\nu,w)\int_w^a e^{-y(\frac{\mu+\lambda_1}{b})}dyd\nu dw \\ &= \frac{e^{\lambda_1 a}}{s+\frac{\mu+\lambda_1}{b}}\int_0^a (e^{-ws}-e^{\frac{\mu+\lambda_1}{b}(w-a)-as})\int_a^{\frac{w}{b}+a}\pi(\nu,w)e^{-\lambda_1\nu}d\nu dw \\ &= \frac{e^{\lambda_1 a}}{s+\frac{\mu+\lambda_1}{b}}\Big(\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s) - e^{-(s+\frac{\mu+\lambda_1}{b})a}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big),\end{aligned} \tag{A.12}$$

$$\begin{aligned}
&\int_0^a e^{-sy}\int_y^a\int_a^{\frac{w}{b}+a}\pi(v,w)e^{-\lambda_1(v-a)-\frac{\lambda_2}{b}(w-y)}dvdwdy\\
&=e^{\lambda_1 a}\int_0^a e^{-w(\frac{\lambda_2}{b})}\int_a^{\frac{w}{b}+a}e^{-\lambda_1 v}\pi(v,w)\int_0^w e^{y(\frac{\lambda_2}{b}-s)}dydvdw\\
&=\frac{e^{\lambda_1 a}}{\frac{\lambda_2}{b}-s}\int_0^a(e^{-ws}-e^{-\frac{\lambda_2}{b}w})\int_a^{\frac{w}{b}+a}\pi(v,w)e^{-\lambda_1 v}dvdw\\
&=+\frac{e^{\lambda_1 a}}{\frac{\lambda_2}{b}-s}(\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s)-\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,\frac{\lambda_2}{b})).
\end{aligned}\tag{A.13}$$

Therefore, we obtain:

$$\begin{aligned}
\frac{b(1+\rho)}{\lambda_2}\tilde{\pi}_Y(s)=&\frac{1-e^{-(s+\frac{\mu+\lambda_1}{b})a}}{s+\frac{\mu+\lambda_1}{b}}(\pi(a,0)+\tilde{\pi}_Y(s))+\frac{1-e^{-(s-\frac{\lambda_2}{b})a}}{s-\frac{\lambda_2}{b}}\Big(\tilde{\pi}_Y(\frac{\lambda_2}{b})\\
&-\tilde{\pi}_Y(s)+e^{\lambda_1 a}(\tilde{\pi}_X(\lambda_1+\frac{\lambda_2}{b})+\tilde{\pi}_{X,Y;Y>a}(\lambda_1,\frac{\lambda_2}{b}))\Big)+\frac{e^{\lambda_1 a}}{s+\frac{\mu+\lambda_1}{b}}\\
&\times\Big(\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s)-e^{-(s+\frac{\mu+\lambda_1}{b})a}\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big)\\
&+\frac{e^{\lambda_1 a}}{\frac{\lambda_2}{b}-s}(\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s)-\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,\frac{\lambda_2}{b})).
\end{aligned}\tag{A.14}$$

And this completes the proof.

2. **Proof for Lemma 5.4.3**: *The LST of the stationary distribution for state set* $(x,x);x\ge a$ *is according to:*

$$\begin{aligned}
\tilde{\pi}_X(s)=C^{-1}\Big[&\frac{e^{-(s+\frac{\mu+\lambda_1^A}{b})a}}{s+\frac{\mu+\lambda_1^A}{b}}\Big(e^{-\lambda_1 a}(\pi(a,0)+\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b}))+\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big)\\
&+\Big(\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}-\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big)\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s-\lambda_1)\\
&+\frac{e^{-(s-(\lambda_1+\frac{\lambda_2}{b}))a}}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big(\tilde{\pi}_X(\lambda_1+\frac{\lambda_2}{b})+\tilde{\pi}_{X,Y;Y>a}(\lambda_1,\frac{\lambda_2}{b})\Big)\Big],
\end{aligned}\tag{A.15}$$

where $C(s)=\frac{b(1+\rho)}{\lambda_1 b+\lambda_2}-\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}+\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}$.

Proof We re-write the equation (5.27); doing so we obtain,

$$\begin{aligned}\frac{b(1+\rho)}{\lambda_1 b+\lambda_2}\pi(x,x) &= \pi(a,0)e^{-\frac{\mu+\lambda_1^A}{b}x-\lambda_1 a}+e^{-\frac{\mu+\lambda_1^A}{b}x-\lambda_1 a}\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b})+\int_a^x \pi(\nu,\nu)\\ &\times e^{-\frac{\mu+\lambda_1^A}{b}(x-\nu)}d\nu+\int_x^\infty \pi(\nu,\nu)e^{-(\lambda_1+\frac{\lambda_2}{b})(\nu-x)}d\nu+\int_a^x\int_w^{\frac{w}{b}+a}\\ &\times \pi(\nu,w)e^{-\lambda_1(\nu-w)-\frac{\mu+\lambda_1^A}{b}(x-w)}d\nu dw+\int_x^\infty\int_w^{\frac{w}{b}+a}\pi(\nu,w)\\ &\times e^{-\lambda_1(\nu-w)-(\lambda_1+\frac{\lambda_2}{b})(w-x)}d\nu dw+\int_0^a\int_a^{\frac{w}{b}+a}\pi(\nu,w)\\ &\times e^{-\lambda_1(\nu-w)-\frac{\mu+\lambda_1^A}{b}(x-w)}d\nu dw. \end{aligned} \tag{A.16}$$

Next, we take a Laplace transform of the right and left hand side, and simplify the integrations according to the following derivations;

$$\int_a^\infty e^{-sx}\int_a^x \pi(\nu,\nu)e^{-\frac{\mu+\lambda_1^A}{b}(x-\nu)}d\nu dx=\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}\tilde{\pi}_X(s), \tag{A.17}$$

$$\int_a^\infty e^{-sx}\int_x^\infty \pi(\nu,\nu)e^{-(\lambda_1+\frac{\lambda_2}{b})(\nu-x)}d\nu dx=\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}(e^{-(s-(\lambda_1+\frac{\lambda_2}{b}))a}\tilde{\pi}_X(\lambda_1+\frac{\lambda_2}{b})-\tilde{\pi}_X(s)). \tag{A.18}$$

Therefore, we will have:

$$\begin{aligned}\frac{b(1+\rho)}{\lambda_1 b+\lambda_2}\tilde{\pi}_X(s) &= \frac{e^{-(s+\frac{\mu+\lambda_1^A}{b})a}}{s+\frac{\mu+\lambda_1^A}{b}}\Big(e^{-\lambda_1 a}(\pi(a,0)+\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b}))+\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big)\\ &+\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}\Big(\tilde{\pi}_X(s)+\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s-\lambda_1)\Big)\\ &-\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big(\tilde{\pi}_X(s)+\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s-\lambda_1)\Big)\\ &+\frac{e^{-(s-(\lambda_1+\frac{\lambda_2}{b}))a}}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big(\tilde{\pi}_X(\lambda_1+\frac{\lambda_2}{b})+\tilde{\pi}_{X,Y;Y>a}(\lambda_1,\frac{\lambda_2}{b})\Big).\end{aligned}$$

And this completes the proof.

3. **Proof for Lemma 5.4.4**: *The LST of the stationary distribution for state set* $(x,y); y<a$ *is according to:*

$$\tilde{\pi}_{X,Y;Y\leq a}(s_1,s_2)=\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}e^{-as_1}\Big[\frac{1-e^{-a(s_2+\frac{\mu+s_1}{b})}}{(s_2+\frac{\mu+s_1}{b})}-\frac{1-e^{-a(s_2+\frac{\mu+\lambda_1}{b})}}{(s_2+\frac{\mu+\lambda_1}{b})}\Big]$$
$$\times((1-\rho)+\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b}))+\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}e^{-a(s_1-\lambda_1)}\Big[\frac{e^{-a(s_2+\frac{\mu+s_1}{b})}}{-(s_2+\frac{\mu+s_1}{b})}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1,-\frac{\mu+\lambda_1}{b})$$
$$+\frac{1}{(s_2+\frac{\mu+s_1}{b})}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1,s_2+\frac{s_1-\lambda_1}{b})+\frac{e^{-a(s_2+\frac{\mu+\lambda_1}{b})}}{(s_2+\frac{\mu+\lambda_1}{b})}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1,-\frac{\mu+\lambda_1}{b})$$
$$-\frac{1}{(s_2+\frac{\mu+\lambda_1}{b})}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1,s_2)\Big]. \tag{A.19}$$

Similarly for the state set $(x,y);y\geq a$ *we will have:*

$$\tilde{\pi}_{X,Y;Y>a}(s_1,s_2)=\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}\Big[e^{-a(s_1-\lambda_1)}(\frac{e^{-a(s_2+\frac{\mu+s_1}{b})}}{(s_2+\frac{\mu+s_1}{b})})-(\frac{e^{-a(s_1+s_2+\frac{\mu+\lambda_1^A}{b})}}{(s_1+s_2+\frac{\mu+\lambda_1^A}{b})})\Big]$$
$$\times\Big(e^{-\lambda_1 a}(\pi(a,0)+\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b}))+\tilde{\pi}_X(-\frac{\mu+\lambda_1^A}{b})+\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\Big)$$
$$+\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}\Big[\frac{e^{-a(s_1-\lambda_1)}}{(s_2+\frac{\mu+s_1}{b})}\tilde{\pi}_{X,Y;Y>a}(\lambda_1,\frac{s_1-\lambda_1}{b}+s_2)-\frac{1}{(s_1+s_2+\frac{\mu+\lambda_1^A}{b})}$$
$$\times\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s_1+s_2-\lambda_1)\Big]. \tag{A.20}$$

Proof We re-write the equation (5.28) and take the Laplace transform of both sides of the equation to obtain,

$$\pi(x,y;y<a)=\pi(a,0)\frac{\mu}{b}\lambda_1 e^{-(\mu+\lambda_1)\frac{y}{b}}e^{\lambda_1(x-a)}+\int_0^a\pi(a,w)\frac{\mu}{b}\lambda_1 e^{-(\mu+\lambda_1)\frac{y-w}{b}}e^{\lambda_1(x-a)}dw$$
$$+\int_0^y\int_a^{\frac{w}{b}+a}\pi(\nu,w)\frac{\mu}{b}e^{-\mu\frac{y-w}{b}}\lambda_1 e^{-\lambda_1(\frac{y-w}{b}+\nu-x)}d\nu dw. \tag{A.21}$$

Next, we take the Laplace transform of the right and left hand side, and simplify the integrations according to the following derivations;

$$\int_a^\infty e^{-sx}\int_0^a\int_a^{\frac{w}{b}+a}\pi(\nu,w)e^{-(\lambda_1)(\nu-w)-\frac{\mu+\lambda_1^A}{b}(x-w)}d\nu dw dx=\frac{e^{-(s+\frac{\mu+\lambda_1^A}{b})}}{s+(\frac{\mu+\lambda_1^A}{b})}\tilde{\pi}_{X,Y;Y\leq a}(\lambda_1,-\frac{\mu+\lambda_1}{b}), \tag{A.22}$$

$$\int_a^\infty e^{-sx}\int_a^x\int_w^{\frac{w}{b}+a}\pi(\nu,w)e^{-(\lambda_1)(\nu-w)-\frac{\mu+\lambda_1^A}{b}(x-w)}d\nu dw dx=\frac{1}{s+(\frac{\mu+\lambda_1^A}{b})}\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s-\lambda_1), \tag{A.23}$$

$$\int_a^\infty e^{-sx}\int_x^\infty\int_w^{\frac{w}{b}+a}\pi(v,w)e^{-(\lambda_1)(v-w)-(\lambda_1+\frac{\lambda_2}{b})(w-x)}dvdwdx$$
$$=\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}(e^{-(s-(\lambda_1+\frac{\lambda_2}{b}))a}\tilde{\pi}_{X,Y;Y>a}(\lambda_1,\frac{\lambda_2}{b})-\tilde{\pi}_{X,Y;Y>a}(\lambda_1,s-\lambda_1)), \tag{A.24}$$

Therefore, we will have:

$$\begin{aligned}\tilde{\pi}_{X,Y;Y\le a}(s_1,s_2)&=\pi(a,0)\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}e^{-as_1}\Big[h(s_1)-g(s_1)-h(\lambda_1)+g(\lambda_1)\Big]\\&+\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}e^{-as_1}\Big[h(s_1)-g(s_1)-h(\lambda_1)+g(\lambda_1)\Big]\tilde{\pi}_Y(-\frac{\mu+\lambda_1}{b})\\&+\frac{\mu}{b}\frac{\lambda_1}{\lambda_1-s_1}e^{a(\lambda_1-s_1)}\Big[h(s_1)\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s_2+\frac{s_1-\lambda_1}{b})-g(s_1)\\&\quad\times\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})+g(\lambda_1)\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,-\frac{\mu+\lambda_1}{b})\\&-h(\lambda_1)\tilde{\pi}_{X,Y;Y\le a}(\lambda_1,s_2),\end{aligned}$$

when $g(s_1)=\frac{e^{-a(s_2+\frac{\mu+s_1}{b})}}{(s_2+\frac{\mu+s_1}{b})}$ and $h(s_1)=\frac{1}{(s_2+\frac{\mu+s_1}{b})}$.

Finally, applying the same procedure to equation (5.29) will give us (5.34).

A.2.1 More details related to Section 5.5

$$\begin{aligned}\tilde{\pi}_X(s)=C(s)^{-1}\Big[&\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}\pi(0,0)+\Big(\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}-\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big)\tilde{\pi}_{X,Y}(\lambda_1,s-\lambda_1)\\&+\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}\Big(\tilde{\pi}_X(\lambda_1+\frac{\lambda_2}{b})+\tilde{\pi}_{X,Y}(\lambda_1,\frac{\lambda_2}{b})\Big)\Big]\end{aligned} \tag{A.25}$$

where $C(s)=\frac{b(1+\rho)}{\lambda_1 b+\lambda_2}-\frac{1}{s+\frac{\mu+\lambda_1^A}{b}}+\frac{1}{s-(\lambda_1+\frac{\lambda_2}{b})}$.

Curriculum Vitae

Name:	Maryam Mojalal
Post-Secondary Education and Degrees:	Allameh Tabataba'i University Tehran, Iran 2008 - 2010 M.Sc. (Mathematical statistics) University of Western Ontario London, ON 2014 - 2018 Ph.D. (Statistics)
Related Work Experience:	Graduate Teaching Assistant The University of Western Ontario 2014 - 2018 Statistical Cunsultant The University of Western Ontario 2015 - 2018 Data (sales) Analyst MTN-Irancell Telecom Co. 2011 - 2012

Publications:

The Lowest Priority Waiting Time Distribution in the Affine and the Delayed Accumulating Priority Queues (submitted)

Discrete time Markov chain algorithm for short time predictions in an Emergency Department (to be submitted)