Waiting-Line Problems with Priority Assignment, and its Application on Hospital Emergency Department Wait-Time

by

Hsing-Ming Chang

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfilment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Statistics

University of Manitoba

Winnipeg

Abstract

The aim of this thesis is to first give a brief review of waiting line problems which often is a subject related to queueing theory. Simple counting processes such as the Poisson process and the duration of service time of each customer being exponentially distributed are often taught in a undergraduate or graduate stochastic process course. In this thesis, we will continue discussing such waiting line problems with priority assignment on each customer. This type of queueing processes are called priority queueing models.

Patients requiring ER service are triaged and the order of providing service to patients more than often reflects early symptoms and complaints than final diagnoses. Triage systems used in hospitals vary from country to country and region to region. However, the goal of using a triage system is to ensure that the sickest patients are seen first. Such wait line system is much comparable to a priority queueing system in our study. The finite Markov chain imbedding technique is very effective in obtaining the waiting time distribution of runs and patterns. Applying this technique, we are able to obtain the probability distribution of customer wait time of priority queues. The results of this research can be applied directly when studying patient wait time of emergency medical service. Lengthy ER wait time issue often is studied from the view of limited spacing and complications in hospital administration and allocation of resources. In this thesis, we would like to study priority queueing systems by mathematical and probabilistic modeling.

Acknowledgments

I give honors and thanks to the Lord who is always there for me wherever I go. “In nothing be anxious, but in everything, by prayer and petition with thanksgiving, let your requests be made known to God;” (Philippians 4:6).

I am deeply grateful to my direct supervisor, Dr. James C. Fu, whose support, teaching, suggestions and encouragement helped me all the time in learning and writing of this thesis. I thank you for your constant encouragement and curricular and extracurricular guidance. You truly make my student life at the University of Manitoba enjoyable.

I am also very thankful to the members of my Ph.D. committee, Dr. S. Mandal, Dr. P. N. Shivakumar and Dr. L. Wang (in alphabetical order by last name) for monitoring my work and taking effort in reading and providing valuable suggestions and comments. I am also very grateful to Dr. Jiann-Ming Wu agreeing to be the external examiner on the committee and his useful comments about my research topic. Much appreciation goes to Dr. Diana McMillan for being the chair of my oral defense and keeping the entire examination process smooth. I would also like to offer personal thanks to Margaret Smith for taking care of all our departmental administrative work and offering her assistance when needed.

At last, I will like to give thanks to the Changhua Christian Hospital Erlin Branch in Taiwan for providing their data. Much appreciation goes to the entire nursing team and staff for their contribution to my better understanding of the flow and operation

of emergency medical service during my visit at the branch. Special thanks to Miss Shu-Yi Wu as my primary contact person at the Changhua Christian Hospital Erlin Branch in Taiwan. Much of your assistance is deeply appreciated.

To my parents,

To grandma, you always want me to be

Contents

List of Tables

List of Figures

Chapter 1

Introduction

ER Wait-Time Stories "Despite growing public pressure and a request from Manitoba's health minister, the Winnipeg health authority refused Tuesday to release an internal investigation into the death of Brian Sinclair last September." Winnipeg Free Press reported in February of 2009 regarding the tragedy happened in September of 2008. Brian Sinclair, a 45 year-old First Nation's double leg amputee died at the Winnipeg Health Sciences Centre (WHSC) while waiting for more than 30 hours for care. The cause of his death was later determined due to a catheter blockage and a bladder infection which were entirely preventable. Dr. Brock Wright, the head of the Winnipeg Regional Health Authority (WRHA), confirmed by examining security tapes days after the tragedy that Sinclair had spoken to an employee at the triage desk within an hour of his arrival, to housekeeping staff and a security guard during his time in the emergency department, but was never properly triaged, registered or assessed by any medical staff. It was also reported that short of staff was not a problem at the WHSC, for during the day and early evening there were at least three health workers helping to process emergent patients.

"'**A disgrace'** Senior waiting in ER at Montfort was ignored for nine hours as fellow patients brought him water, blankets", reported by the Ottawa Citizen in September of 2008. Yatendra Varshni, the 76 year-old professor emeritus of the University of Ottawa, was sent to the hospital by ambulance around 3 p.m. on September 26, 2008 and later was determined that he suffered from rheumatoid arthritis. Mr. Varshni was admitted around 12:30 a.m. the next morning, waiting for more than 9 hours in the hospital's ER while other fellow patients brought him water and blankets, and later testify to CTV Ottawa that the on-duty nurse ignored him.

A quick search on the Internet returns many news and reports on topics of hospital wait times. Though many studies focus on hospital staff management and policy reformation, we wish to the study emergency department wait time from the perspective of statistical and probabilistic modeling using the finite Markov chain imbedding (FMCI) technique.

1.1 Background of Queueing

A queueing system, in its simplest description, consists of *customers* arriving at some random times to a waiting line, each waits for some random amount of time before receiving service, and the service-time of each *customer* is random according to some probability distribution F. *Customers* depart from the system after being served. The word *customers* is used as a generic term and it may refer to, for example, airplanes arriving to an airport, shoppers in a grocery store waiting in line to checkout their goods purchased, incoming calls in a telecommunication system waiting to be transmitted, or tasks in a computer system waiting to be executed by the processing unit, and etc. In another setting, locations requires, stochastically in time, service can also be seen as customers, ie. sites on fire waiting for fire fighters to arrive and put out fire; crime scenes waiting for police force to arrive, etc. In applications, priority queues can be used to model an objective function as to reduce measures such as system cost or outcome casualty.

Agner Krarup Erlang, a Danish engineer at the Copenhagen Telephone Company in Denmark, published his first paper *The theory of probabilities and telephone conversations* in 1909, showed that incoming calls in telephone traffic can be characterized by the Poisson distribution. At the time, J. Jensen (of Jensen's inequality) was the chief engineer at the company. Later, Erlang published another paper *Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges* in 1917 that includes the famous Erlang loss and waiting time formulas. The two and his many other papers were translated into English, French and German. His concept of "statistical equilibrium" were used to justify some ergodic results and study system behavior. Although rigorous proofs were not presented, Erlang laid the foundations for modern queueing theory.

In general, a formal mathematical modeling of a queueing system requires one to specify assumptions made about (i) the input process, (ii) the queueing discipline, and (iii) the service mechanism.

The input process describes the way customers arrive to a system. In this section only, let t_k, for $k \geq 1$, denote the arrival time of the kth customer, $t_0 < t_1 < t_2 < t_k < \cdots$. With loss of generality, let $t_0 = 0$ be the initial time when we began to observe the input process. Let $T_k > 0$, $T_k = t_k - t_{k-1}$, be the inter-arrival time between the $(k-1)$th and kth customer. The usual assumption is that T_k, $k = 1, 2, \cdots$, be a sequence of independent and identically distributed random variables with some distribution $G(x) = P(T_k < x)$. $G(x)$ is referred to as the distribution of inter-arrival time and often is assumed to have the form of an exponential distribution, $G(x) = 1 - exp\{-\lambda x\}$, having mean inter-arrival time $1/\lambda$. Hence, λ can be regarded as the mean arrival rate. Under such setting, it can be shown that the counting process $N(t)$, the number of arrivals to time t, follows the Poisson distribution having density

$$\Pr\{N(t) = n\} = \frac{(\lambda t)^n}{n!}\, exp\{-\lambda t\},$$

see [18].

Jaiswal [18] in his book *Priority Queues* distinguishes the *source* from which customers emanate either being finite for infinite. The distribution of inter-arrival time differs slightly depending on the source being finite or infinite, and the definition of the inter-arrival times. For our purposes, we assume that the source is always infinite in this thesis.

A queueing discipline consists of rules by which customers are ordered in line for service. The simplest rule in a single-server system is the *first come, first serve* (FCFS) discipline by which customers receive service in the order of their arrival. Other service policies, such as *random-service*, *last come*, *first serve*, *batch-service*, *service with vacations*, *priority service*, *deadline-ordered* service, and many others are possible to be employed depending on operational requirement and system efficiency to be achieved.

The service mechanism of a queueing system describes the output end of the system. The output process contains information on the number of servers and the service-time distribution. A system can have one or more servers, or sometimes called service channels. Systems with only one channel are called single-server queues. Systems with more than one channel are called multi-server queues. Let $S_k > 0$ denote the service time of the kth arrived customer. The usual assumption is that S_k, $k = 1, 2, \cdots$, be a sequence of independent and identically distributed random variables with a distribution $F(x) = P(S_k < x)$. $F(x)$ is commonly being referred to as the service-time distribution with mean service time

$$\frac{1}{\mu} = \int_0^\infty (1 - F(x))dx = \int_0^\infty xdF(x) \tag{1.1.1}$$

where $\frac{d}{dx}F(x)$ is assumed to exist.

Here, μ can be interpreted as the expected number of customer departures from the service channel per unit time when $F(x)$ is an exponential distribution. If W

denotes the expected time a customer spent waiting in line before receiving service, then the expected number of customers waiting in line

$$L = \lambda \times W \tag{1.1.2}$$

which depends only on the "long run mass flow balance relations" as described in Taylor and Karlin [32] page 543. This equation is of great importance in queueing theory in evaluating the performance of queueing systems in many application, since it directly relates two of the most important factors which are the average queue size L and the average customer waiting time W. For example, in an emergency hospital, patient satisfaction often is related to the amount of time they need to wait before receiving treatments. Therefore, waiting time reduction may be of great importance in a hospital queueing system, see Anderson, Black, Dun and etc. [30] and Spaite, Bartholomeaux, Guisto and etc. [3] for example.

David G. Kendall in 1953 introduced the $A/B/C$ Kendall's notation to simplify the way of describing and classifying queueing systems. When one wishes to describe a queueing system, the lengthy procedure of having to specify the input process, the queueing discipline, and the service mechanism becomes compact and standard. The first component of $A/B/C$ describes the input or arrival process. Sometimes the first letter also is used to denote the probability distribution of inter-arrival time length. The first component of $A/B/C$ describes the output process, or sometimes it is used to denote the probability distribution of service time of each customer. The third letter denotes the number of servers in the system. An example of using such notation system is a $M/M/1$ queue where of arrival process the inter-arrival time is assumed to follow an exponential distribution (the letter 'M' may be used to remind the Markovian property of exponential distributions), the service-time distribution is again exponential, and there is only one server in the described system. For a few other commonly studied queueing systems such as $E_k/G/1$ and $GI/M/s$ etc., please see Kendall [20].

Under the assumption that customers arrive to the system as a Poisson process with rate λ, there is only a single server, and the mean service-time of priority class i customers be finite and without specifying the form of service-time distribution, Cobham([7, 8]) was the first to consider the waiting line problem with service priorities assigned to customers and found an expression for the expectation of waiting time. In [7], essentially the model considered is now what we called the non-preemptive priority queue in which only one customer at a time can be in service. Further results by Holley [16], Kesten and Runnenberg [21], Aczel [1] and Miller [28], some under different settings in the output process, generated many applications in the analysis of priority queues.

1.2 Early Results of Priority Queuing Model

After reviewing the earliest papers on priority queues, we will use Cobham's [7] definitions and notation to understand some of the early results and how they were derived. Suppose that each customer arriving at a single-channel waiting line can be categorized into one of r classes (r is a finite positive integer) corresponding to r independent Poisson processes with rates $\lambda_1, \lambda_2, \ldots, \lambda_r$, respectively.

A customer U of priority p (denoted by U_p from now on), $1 \leq p \leq r$ (1 indicates the highest priority and r the lowest), enters a system and is moved ahead of all customers with priority level $k > p$ larger than p and behind all with priority level $k < p$. If there already exist customers with priority level p in the waiting line, customers of the same level will be served in the first-come, first serve order within class p. Let $F_p(t)$ be an arbitrary customer service-time distribution of unit U_p. Then,

$$F(t) = \sum_{p=1}^{r} \frac{\lambda_p}{\lambda} F_p(t) \tag{1.2.1}$$

was defined by some as the combined customer service-time distribution.

Recall from equation (1.1.1), we then let

$$\frac{1}{\mu_p} = \int_0^\infty t \, dF_p(t). \tag{1.2.2}$$

denote the expected service time of a priority p customer, in a $M/M/1$ priority queue, independent of other priority customers.

Now, for a simple and clear presentation of the derivation of some of the results in Cobham [7], we will continue but with the notations used in Holley's [16] paper. Suppose that a customer U_p enters the waiting line at time t_0 and receives service at time t_1. Thus the length of time U_p waited in line is $T = t_1 - t_0$. Suppose at time t_0, there is n_0 customer in service ($n_0 = 0$ or 1), and there are n_k customers of priority k ($k = 1, 2, \ldots, p$) in the line ahead of U_p. To complete the service of the customer already at the counter will take time T_0 and to serve the n_k customers of priority k currently in line ahead of U_p will take time T_k. The variable T is random and during the entire time T, customers of priority level less than p will continue to enter the system and take places in the line ahead of U_p. Suppose n'_k customers of priority k (for $k = 1, 2, \ldots, p-1$) entered during T and it takes T'_k amount of time to service them. Under the assumption that there is a single server and customers receive their service in succession without time gaps in between, the length of T must equal the sum of T_0, the p quantities T_k and $p-1$ quantities T'_k. Taking expectation of T, as equation (6) in Holley [16], we have the wait time model

$$E[T] = \sum_{k=1}^{p-1} E[T'_k] + \sum_{k=1}^{p} E[T_k] + E[T_0] \tag{1.2.3}$$

If we were to use Holley's [16] notation, set

$$W_0 = \int_0^\infty \frac{1}{2}\, \lambda t^2 \, dF(t) = E[T_0]$$

which in Cobham [7] is the expected service time of the customer at the counter at time t_0 the instant U_p arrives at the waiting line. The intuition for the definition of the W_0 is unclear to us. Let W_k denote the expected waiting time for a customer of priority k ($k = 1, 2, \ldots, p$). It is clear that $W_p = E[T]$. In [7], W_p was derived to be

$$W_p = \frac{W_0}{\left(1 - \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k}\right)\left(1 - \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k}\right)}$$

We will attempt to show the derivation of the above expression in the following.

The expected time $E[T_k]$ required for the service of n_k customers of priority k in line at time t_0 is the expected service time $\frac{1}{\mu_k}$ multiplied by the expected value $E[n_k] = \lambda_k W_k$, by (1.1.2). The expected time $E[T_k']$ required for the service of n_k' customers of priority k $(1 \leq k \leq p-1)$ entering the waiting line after the arrival of U_p, similarly, is the expected service time $\frac{1}{\mu_k}$ multiplied by the expected value $E[n_k'] = \lambda_k E[T]$ where $E[T] = W_p$. Thus, equation (6) in Holley [16] can be written as

$$W_p = \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k} W_p + \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k} W_k + W_0 \tag{1.2.4}$$

By re-arranging the above equation to isolate W_p, it follows that

$$\begin{aligned} W_p &= \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k} W_p + \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k} W_k + W_0 \\ \left(1 - \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k}\right) W_p &= \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k} W_k + W_0 \\ W_p &= \frac{\sum_{k=1}^{p} \frac{\lambda_k}{\mu_k} W_k + W_0}{\left(1 - \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k}\right)} \end{aligned}$$

Alternatively, from (1.2.4), W_p can also be expressed as

$$\begin{aligned} W_p &= \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k} W_p + \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k} W_k + W_0 \\ W_p &= W_p \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k} + \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k} W_k + W_0 \\ \left(1 - \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k}\right) W_p &= \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k} W_k + W_0 \\ W_p &= \frac{\sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k} W_k + W_0}{\left(1 - \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k}\right)} \end{aligned}$$

with $\frac{\lambda_0}{\mu_0} = 0$.

In this form, we can show $W_1 = \frac{\mu_1}{\mu_1 - \lambda_1} W_0$. Dropping the index 1 gives the expected waiting time W of the classical single-server Poisson process with a mean arrival rate λ and service-time distribution $F(t)$, of which the mean service rate (rate of customer departure) is μ as defined in (1.1.1). By (1.2.5) and substituting $W_1 = \frac{\mu_1}{\mu_1 - \lambda_1} W_0$, W_2 can be solved to be

$$\begin{aligned} W_2 = & \frac{\frac{\lambda_1}{\mu_1} W_1 + W_0}{\left(1 - \sum_{k=1}^{2} \frac{\lambda_k}{\mu_k}\right)} \\ = & \frac{\frac{\lambda_1}{\mu_1 - \lambda_1} W_0 + W_0}{\left(1 - \sum_{k=1}^{2} \frac{\lambda_k}{\mu_k}\right)} \\ = & \frac{W_0}{\left(1 - \frac{\lambda_1}{\mu_1}\right)\left(1 - \sum_{k=1}^{2} \frac{\lambda_k}{\mu_k}\right)} \end{aligned}$$

By induction on p, for $p \geq 2$, W_p can be solved to be

$$W_p = \frac{W_0}{\left(1 - \sum_{k=1}^{p-1} \frac{\lambda_k}{\mu_k}\right)\left(1 - \sum_{k=1}^{p} \frac{\lambda_k}{\mu_k}\right)}$$

which is equation (3) in Cobham [7]. Under very specific settings, we see that only the expectations can be derived but not the probability distribution of the random variables W_p, $p = 1, 2, \ldots, r$.

Kesten and Runnenburg [21] in 1957 gave a rigorous proof in a more detailed account of the single service-counter situation described in Cobham's [7] setting by analysis of continuous time Markov chain. They proved that, independent of the initial state of the queue, the probability distribution function ($H_p(t)$ in their notation) of waiting time W_p of a customer assigned arbitrarily priority $p \in \{1, \ldots, r\}$ exists in the very restriction that only when the system is in stationary and non-saturation state. By non-saturation, it means $\sum_{i=1}^{r} \frac{\lambda_i}{\mu_i} < 1$.

If we were to treat the priority queue system as a first-come, first-serve combined queue system, meaning there is no priority assignment to units in a single service line, then Aczel [1] showed that the expected queue length, name it L_1, can be expressed as

$$L_1 = \frac{W_0 \sum_{k=1}^{r} \lambda_k}{1 - \sum_{k=1}^{r} \frac{\lambda_k}{\mu_k}}$$

For the same priority queue system, Cobham [7] gave the expected queue length, name it L_2, to be

$$L_2 = \sum_{j=1}^{r} \frac{\lambda_j W_0}{\left(1 - \sum_{k=1}^{j-1} \frac{\lambda_k}{\mu_k}\right)\left(1 - \sum_{k=1}^{j} \frac{\lambda_k}{\mu_k}\right)}$$

By some algebraic work, Aczel in [1] showed that

$$L_2 - L_1 = \sum_{j=2}^{r} \left[\frac{\lambda_j W_0}{\left(1 - \sum_{k=1}^{j-1} \frac{\lambda_k}{\mu_k}\right)\left(1 - \sum_{k=1}^{j} \frac{\lambda_k}{\mu_k}\right)} \sum_{i=1}^{j-1} \lambda_i \left(\frac{1}{\mu_i} - \frac{1}{\mu_j}\right) \right]$$

From this expression, we can see that if we were able to choose $\left(\frac{1}{\mu_i} - \frac{1}{\mu_j}\right) \leq 0$ for $i \leq j$, to assign higher priority to serve first the customers with shorter expected service time, then the expected queue length can be reduced in a queueing system with priority assignment compared to one with the FCFS rule.

Early papers studying these types of queue system were done by solving mathematical differential equations of deterministic models. Then method of induction were used to obtain the expectation of wait times.

The idea of using an imbedded Markov chain to study the behavior of a queueing system was also first roused by Kendall ([19],[20]). Miller [28] attacked waiting-line problem of $M/G/1$ preemptive-resume priority queues using the Markov chain imbedding approach. Both authors examine systems at epochs of customer departures. Stanford [31] studied waiting time and inter-departure time of $\sum M_i + GI/G_i/1$ single-server multi-class priority queues under non-preemptive and preemptive resume disciplines. In [31], a comprehensive review of literatures studying priority queues is given since Cobham ([7],[8]) in the mid nineteen fifties. Solutions of waiting-time distribution classically are provided as LSTs in earlier papers, compare to more recent approach by Alfa [2], Wagner [36], and Ramachandran [29] where matrix-analytical methods are employed. LSTs often are still used in matrix-analytical methods when the inter-arrival time and service-time distributions are not exponential. Wagner [37] and Wagner et al. [38] studied the stationary and waiting distributions of finite-capacity non-preemptive priority queues. In [37], Wagner considered $M/M/s$ multi-class priority with customer service times being identically and exponentially distributed. In [38], $M/M/1$ two-class priority model are considered with class dependent exponential service times. In both papers, stationary distributions are analysed by solving system of Chapman-Kolmogorov equations. By applying matrix-analytic methods, meaningful conditional waiting-time distributions of each priority class are obtained in LSTs. A recursive algorithm to compute the mean waiting time is derived by first-passage-time analysis in [38].

In more recent years, Bedford and Zeephongsekul [5] adopted the approach of [37] and [38] to study the stationary distribution of $M/M/1$ two-class preemptive priority dual queue model with finite capacity. Zeephongsekul and Bedford later in [40] carried out waiting time analysis complementing the work of [5]. Li and Zhao [24] studied the tail asymptotics of the stationary distribution of $M/M/1$ two-class preemptive priority queues by studying the generating function of the joint stationary distribution for the number of customers in both classes. Xie et al. [39] studied the tail asymptotics of the stationary distribution of $M/M/s$ multi-class preemptive priority queues by matrix-analytic methods. Zhang and Shi [41] studied $M/M/1$ two-class preemptive priority queue with infinite queue capacity using the QBD process that its stationary distribution can be exactly computed in principle.

We wish to dissect the priority queueing problem using finite Markov chains. This approach saves us from solving complicated system of equations, if they are solvable, also it allows us to obtain the distribution of wait times easily. Desirable results such as tail probabilities and various moments of the distribution can also be obtained through a relatively simple setup.

1.3 Definitions, Models, Notations and Variables of Interest

1.3.1 Emergency Service Flow

Borrowing from one of the public reports [6] made available by the Canadian Institute of Health Information (CIHI), with some modification we made the following diagram to clearly present the typical emergency medical service flow and some variables of interest later being formulated in our problem.

When patients arrive on their own at an emergency department, they are first triaged by a triage nurse to evaluate the severity of their injury or illness and may be assigned a score according to the 5-level Canadian Triage and Acuity Scale (CTAS).

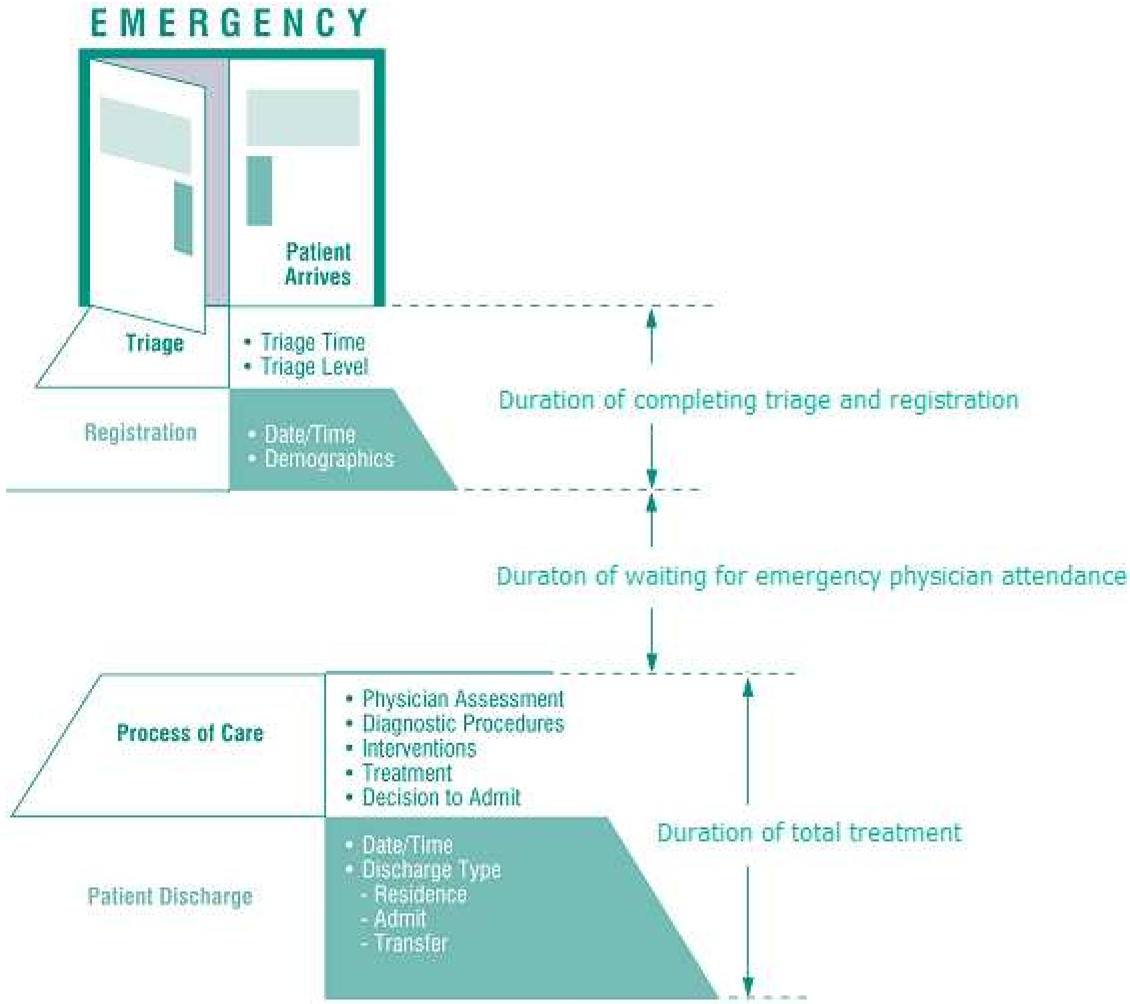

Figure 1.1: Emergency Service Flow Diagram

The CTAS is designed to ensure that patients who require immediate care receive medical attention first. Those with less urgent conditions, such as mild abdominal pain, headache, or conditions related to chronicle problems etc. usually can wait to receive treatment. Then a nurse would input patient information (basic data such as the symptoms observed, vital signs, trauma mechanism and other medical information such as allergies, medications taken and medical attention received prior to arrival etc.) acquired from patients themselves, their accompaniment, or sometimes by estimation. The triage process usually should take no longer than a few minutes.

An arrival (or the accompaniment) then would be asked to register and be identified as a patient requiring care, then be provided with his/her medical record. For patients arrived by ambulance, the registration process might differ, but the severity of patients' injuries or illness still would be assessed. Upon the completion of registration, patients would rest in waiting areas and wait for an emergency physician (EP) to attend. From time to time, nurses would reassess patients' condition. A patient might be placed ahead in queue or require immediate attention if condition deteriorated. For our purpose, we define the time from the completion of triage and registration to the time of initial assessment of a patient by an EP the **ED wait time**.

Starting from the time a patient is triaged to the time the patient departs from the ED, we define this to be the **length of stay**. This length of time includes the waiting time plus the length of time from initial visit of an EP to the patient until patient departure. During such period, the EP may order for additional diagnostic examinations. The EP may or may not attend other patients while waiting for results from orders, then provides interventions and treatments with assisting nursing staff. Finally decisions are made that either patients would be discharged home, admitted to a ward, or transferred to another department or hospital, and be considered as departed from the ED.

1.3.2 Model Description and Notation

We will introduce in this section some of the notations frequently used in Chapter 3 and after in describing priority queues. Some other new ones may be defined along as they are needed. In this thesis, we may use the word *customer(s)* in place of *patient(s)* when describing a queueing model. We mainly consider that:

1. customers are of R classes indicating their service priority class;
2. there can be no more than b customers in the queueing system at any time, including the one(s) receiving service and those waiting in line yet to receive service;
3. no more than c customer(s) are allowed to receive service at any given time.

Further we assume that customers arrive to the system from R independent unscheduled Poisson arrival processes. Customers of class i arrive to the system at a mean arrival rate $\lambda_i > 0$ for $i = 1, \ldots, R$. We use lower indices to indicate higher priorities.

As $\Delta t \to 0$, the probability of having one arrival of class i customer in $[t - \Delta t, t)$ is $\lambda_i \Delta t + o(\Delta t)$, the probability of having more than one arrival in $[t - \Delta t$,t) is $o(\Delta t)$ for any t. We use $o(\Delta t)$ to denote a function of Δt such that $\lim\limits_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$. Jaiswal in [18] distinguishes the *source* from which customers emanate either being finite for infinite. We assume here that the source is infinite.

Let $B(t)$, $0 \leq B(t) \leq b$, be the number of customers occupying the system (including the ones in service and those waiting for service) and $C(t)$, $0 \leq C(t) \leq c < b$, be the number of customers in service at time t. We assume in our model that when $B(t) \leq c$, then service would be provided to all customer in the system keeping no one waiting. If $c < B(t) \leq b$, service would be provided to c customers and keeping $B(t) - c$ waiting in line.

Suppose a customer of priority i, denoted by U_i, arrived to the system during a time interval $[t - \Delta t, t)$. If U_i is admitted to the system, then, departure of U_i is not allowed during the same time interval $[t - \Delta t, t)$. If the system is not empty and U_i has to wait in the queue, U_i would be placed ahead of all customers with service priority scores larger than i and behind all with priority scores of i or smaller starting at time t.

Suppose U_i is to join the queueing system while $c \leq B(t - \Delta t) < b$ customers are in the system. Denote the largest service priority score among those in service by $K_1(t)$ at time t. We consider in this situation both the preemptive repeat-different (PRD) and the non-preemptive (NP) disciplines which have the definitions as in Jaiswal [18] Chapter *III*:

(a) Preemptive repeat-different: If there is no customer of priority score i or smaller waiting for service, U_i would displace a customer in service with the highest service priority score $K_1(t)$. The preempted customer $U_{K_1(t)}$ then waits at the first position in the priority $K_1(t)$ group. When $U_{K_1(t)}$ resumes to service, the remaining service time required is random and is independent of past preemptions and services and has the same exponential service time distribution of class $K_1(t)$.

(b) Non-preemptive: U_i waits behind all with priority scores of i or smaller albeit any customer U_j, $j > i$, is in service at the time of arrival of U_i. The service of U_j, for any $j > i$, continues until completion.

The order of customer service given is based on their priority, and within each priority class FCFS. In this entire thesis, customer service time is assumed to be exponential and is class dependent, and there is only a single server in the system. Priority i customers have an exponential service-time distribution with mean service time $1/\mu_i$, for $i = 1, 2, \ldots, R$.

We will use the notation $M/M/1$-$R/b/c$ to characterize priority queueing system in our model where the the first three characters bear the same information as those introduced by Kendall [20], the parameters R, b and c are as introduced earlier.

For example, in the applications of health-care, an ED typically has only one EP attending several patients. There are several emergency beds for patients who are in treatment or are under observation after initial treatment. At the same time there are more beds for patients who require attention before receiving any treatment and some patients need to wait without a bed available for them. But due to the nature of operations, the models of a priority queue for ED service is not quite the same as a multi-server priority queue studied by others.

Results in [7] can not be readily applied here. In the later chapters, we will discuss more in detail about priority queues and demonstrate the use of the finite Markov chain imbedding technique to obtain the probability distribution of ER wait times.

Chapter 2

Markov Chain

2.1 Basic Definitions

For convenience, we use the notation $X = \{X_t : t \in T\}$ to denote a stochastic process where X_t is a random variable and T is an index set. In this thesis we will treat T a collection of discrete times otherwise indicated. For simplicity, T commonly starts with the value 0 to denote initial time, and often $T = \mathbb{N} \cup \{0\} = \{0, 1, 2, \cdots\}$. But for theoretical purposes and some applications, T can be treated as a continuous set over $[0, \infty)$. In this case, $[0, t) \subset T$.

A Markov process $\{X_t\}$ is a stochastic process with the property that given the value of X_t at some time t, future values X_u for $u > t$ does not depend on past values X_s for $s < t$. We will give more formal terms of the discrete version of a Markov chain in a probabilistic sense that, let $X = \{X_0, X_1, X_2, \cdots\}$ be a discrete time, discrete state space stochastic process with state space Ω. X_t is in Ω for any t and X is said to be a Markov Chain if given states $i_{t+1}, i_t, i_{t-1}, \ldots, i_1, i_0 \in \Omega$ and any time t, we observe

$$\Pr\{X_{t+1} = i_{t+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_{t-1} = i_{t-1}, X_t = i_t\}$$

$$= \Pr\{X_{t+1} = i_{t+1} \mid X_t = i_t\}$$

It is generally accepted to use $X_t = i$ to denote that the process is in state i at time t. Also the conditional probability of $X_{t+1} = j$ given that $X_t = i$, the process enters state j at time $t+1$ from state i at time t, is called the one-step transition probability and is denoted by $p_{ij}^{t,t+1} = P(X_{t+1} = j \mid X_t = i)$. In this thesis, we consider only time-homogeneous discrete-time discrete state space Markov chains, which means that transition probabilities $p_{ij}^{t,t+1} = p_{ij}$ are independent of time t for all $i, j \in \Omega$. Moreover, for any $X_t = i$, the process must enter some state $X_{t+1} = j$ in the finite state space Ω, and clearly p_{ij} must satisfy the following conditions:

$$0 \leq p_{ij} \leq 1 \quad \text{for all } i, j \in \Omega,$$

$$\sum_{j \in \Omega} p_{ij} = 1 \quad \text{for all } i.$$

For computational convenience, the one-step transition probabilities p_{ij} are often arranged in a square matrix

$$P = (p_{ij}) = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.1.1}$$

where p_{ij} is the entry of the ith row and the jth column. P is called a transition probability matrix. A Markov process is completely defined once the transition probability matrix is given or constructed. An initial state probability distribution of a Markov chain is defined as in the following

Definition 2.0.1. Initial probability distribution of a Markov chain X is the probability mass function $\boldsymbol{\xi} = (\xi_i)_{i \in \Omega}$ of the initial state X_{i_0}, ie. $P(X_{i_0} = k) = \xi_k$.

Typically $\boldsymbol{\xi}$ is a row vector. In this thesis, all vectors are in rows unless they are indicated by the symbol $\boldsymbol{'}$ as a superscript on the right, ie. $\boldsymbol{\xi}$ is a row vector and $\boldsymbol{\xi}'$ is a column vector — the transpose of $\boldsymbol{\xi}$. We will show a simple example of calculating the probability of a two-event and three-event Markov process in the following. Suppose there is a sample of four events E_0, E_1, E_2 and E_3 happened sequentially, with E_0 being the state or event where the process has started. We may have the interest to calculate the probability of the process X starts with event zero followed by event one. By the definition of conditional probabilities and properties of a Markov process we have

$$P(X_0 = E_0, X_1 = E_1) = P(E_0, E_1) = P(E_1 \mid E_0)P(E_0) = \xi_0 p_{01}$$

By the same token, the probability of a process of three events must be

$$\begin{aligned} P(E_0, E_1, E_2) &= P(E_2 \mid E_0, E_1)P(E_0, E_1) \\ &= P(E_2 \mid E_1)P(E_1 \mid E_0)P(E_0) \\ &= \xi_0 p_{01} p_{12} \end{aligned}$$

and it can be generalized that

$$P(E_{i_0}, E_{i_1}, \cdots, E_{i_t}) = \xi_{i_0} p_{i_0 i_1} p_{i_1 i_2} \cdots p_{i_{t-2} i_{t-1}} p_{i_{t-1} i_t}$$

If we denote by $p_{ij}^{(m)}$ the probability of a Markov process starting from state E_i and entering state E_j in exactly m steps, and let again ξ_i be the probability of the initial state be E_i of a Markov process, the unconditional probability $p_{*j}^{(m)}$ of the process being in state E_j in exactly m steps can be calculated as

$$p_{*j}^{(m)} = \sum_{i \in \Omega} \xi_i p_{ij}^{(m)}$$

(see Feller [9]).

In matrix analysis it is easy to recognize that with a proper arrangement of $\boldsymbol{\xi}$ and the transition probability matrix P, the probability of the process entering state j given that it started in state i in m steps is the entry in the ith row and jth column of the m-step transition matrix

$$P^{(m)} = P \times P \times \cdots \times P = P^m \tag{2.1.2}$$

and the kth element of the resultant vector of the product of $\boldsymbol{\xi}$ and the m-step transition probability matrix $P^{(m)}$

$$\boldsymbol{\xi} P^{(m)} = \boldsymbol{\xi} \times P \times P \times \cdots \times P = \boldsymbol{\xi} \times P^m \tag{2.1.3}$$

is $P_{*i_k}^{(m)}$, for all $i_k \in \Omega$.

2.2 Introducing the Finite Markov Chain Imbedding

To the best of our searching ability, the first paper published using Markov chains to study queues was by David G. Kendall [19] in 1951. In his paper, Kendall explained in detail a very original idea of looking at "regeneration points" at which customers depart to study the stochastic process describing the fluctuations of queue-size, non-Markovian in general, by making it a Markovian one. The $M/G/1$ queue system was studied by considering the behavior of a certain imbedded Markov chain and obtained the distribution of queue length in statistical equilibrium. The ergodic properties of the system in relation to the value of the relative traffic intensity ρ ($\rho < 1$, $\rho = 1$ and $\rho > 1$) was closely examined by Kendall when there was no special condition made on an maximum queue size (the queue size was only assumed to be countable, the maximum value was not assumed, and therefore the dimension of the probability transition matrices of imbedded Markov chains were stochastic but infinite). This paper received rather many discussions, comments and remarks regarding the analysis of a queueing system by Markov chain imbedding. As Kendall mentioned, the same results were first obtained by Pollaczek [27] in his paper published in German and another paper in Russian by Khintchine [22], each using quite different methods. Pollaczek's paper involved much of a difficult method of analysis. Khintchine however used a different and more simpler approach, but an English translation of the paper was not available at that time. The term *Markov chain imbedding* actually first appeared later in Kendall's [20] paper in 1953.

Fu and Koutras [10] first introduced a unified approach which provides a systematically developed theory and method dealing with some problems of runs and patterns based on the finite Markov chain imbedding (FMCI) technique.

Fu and Lou [13] "provides a rigorous, comprehensive introduction to the finite Markov chain imbedding technique for studying the distributions of runs and patterns from a unified and intuitive viewpoint, away from the lines of traditional combinatorics". The concept of finite Markov chain imbedding is re-visited and the book provides ample of theoretical backgrounds with formal definitions, many important theoretical derivations and concepts were presented by which a systematic approach of using such technique to turn many statistical problems into ones in terms of finding runs and patterns now have solutions. Its utility is illustrated through practical applications to a variety of fields, including the reliability of engineering systems, hypothesis testing, quality control, and continuity measurement in the health care sector. The technique itself continues to receive attentions in the theoretical framework, as well as its direct applications in studying probability distribution and recognition of patterns in DNA sequencing, modeling of longitudinal and survival data and the developing of methods for health care monitoring.

In the paper by Fu [11], the exact and limiting distribution of number of successions of size $2 \leq k \leq n$ in a random permutation generated by n distinct positive integers is obtained by the finite Markov chain approach instead of the traditional combinatorial analysis. Also by Fu [12], a simple formula of the distribution of the scan statistics of window size r for a sequence of n Bernoulli trials or Markov dependent bistate trials was derived. The relative small window size r and short length of sequence n limitation in deriving the formula of the distribution of the scan statistics at the time was overcame. In connection to biology, one result Grégory Nuel demonstrated in [25] was using the FMCI technique to device recursive algorithms to compute the exact CDF or complementary CDF of the local score of one sequence and therefore computing the exact p-value of a statistic for the first time when finding hydrophobic segments in a protein database. These is only a small sample of applications of the finite Markov chain imbedding technique.

In a sketch, the approach of our studying the wait time of patient in ED under the priority queue setting is as the following.

1. Recognize the various status (ie. extract sufficient information such as the number of patients in every priority category, time points of observation, etc.) of an ED and represent them by a collection of patterns.

2. Apply the finite Markov chain technique to construct a Markov process for the study of our interest.

3. In most cases, we are interested in obtaining the distribution of the waiting time, first occurrence or first-passage time in Feller's [9] term, of the above

constructed Markov chain reaching some sub-collection of patterns in the state space given an initial condition.

To avoid duplication and further confusion of readers, we will use the definitions and notation in [13]. Some most important theorems used in later chapters of this thesis will simply be stated at this time, for details and proofs of the theorems and many applications of the Finite Markov Chain Imbedding technique, please see [13].

Let $\Gamma = 0, 1, \ldots, n$ be an index set, and let $\Omega = a_1, a_2, \ldots, a_m$ be a finite state space.

Definition 2.0.2. The non-negative integer-valued random variable $X_n(\Lambda)$ is finite Markov chain imbeddable if:

1. there exists a finite Markov chain $\{Y_t : t \in \Gamma_t\}$ defined on a finite state space Ω with initial probability vector $\boldsymbol{\xi}_0$,
2. there exists a finite partition $\{C_x : x = 0, 1, \ldots, l_t\}$ on the state space Ω, and
3. for every $x = 0, 1, \ldots, l_n$, we have

$$P(X_n(\Lambda) = x) = P(Y_n \in C_x \mid \boldsymbol{\xi}_0).$$

Theorem 2.1. *If $X_n(\Lambda)$ is imbeddable by a time homogeneous finite Markov chain, then*

$$P(X_n(\Lambda) = x) = \boldsymbol{\xi}_0 M^n \mathbf{U}'(C_x), \tag{2.2.1}$$

where $\mathbf{U}'(C_x) = \sum\limits_{r:a_r \in C_x} \boldsymbol{e}_r$, $\boldsymbol{e}_r$ *is a* $1 \times m$ *unit vector corresponding to state* a_r, $\boldsymbol{\xi}_0$ *is the initial probability vector, and* M *is the transition probability matrix of imbedded Markov chain.*

If the imbedded Markov chain is non-homogenous, the above theorem still holds with a modification of equation (2.2.1) to

$$P(X_n(\Lambda) = x) = \boldsymbol{\xi}_0 \left(\prod_{t=1}^{n} M_t \right) \mathbf{U}(C_x), \tag{2.2.2}$$

where $\{M_t\}_{t=1}^{n}$ is the sequence of $m \times m$ transition probability matrices of the imbedded finite Markov chain Y_t defined on the state space Ω with initial probability distribution $\boldsymbol{\xi}_0 = (P(Y_0 = a_1), P(Y_0 = a_2), \ldots, P(Y_0 = a_m))$.

Let M be the $m \times m$ transition probability matrix of the finite Markov chain Y_i defined on the state space Ω with initial probability distribution $\boldsymbol{\xi}_0 = (P(Y_0 = a_1), P(Y_0 = a_2), \ldots, P(Y_0 = a_m))$.

A state $\alpha \in \Omega$ is called an absorbing state if once the Markov process enters state α and never leaves the state again; ie. $p_{\alpha\alpha} \equiv 1$, or $p_{\alpha\beta} \equiv 0$ for any $\beta \neq \alpha \in \Omega$. Let $A = \{\alpha_1, \ldots, \alpha_k\}$ be the set of all absorbing states of a time-homogeneous Markov chain Y_t with transition probability matrix M. With proper arrangement of the state space, M can always be expressed in the form:

$$M = \left(\begin{array}{c|c} N_{(m-k)\times(m-k)} & C_{(m-k)\times k} \\ \hline O_{k\times(m-k)} & I_{k\times k} \end{array} \right) \tag{2.2.3}$$

where m and k $(m > k)$ are the numbers of states in Ω and A, respectively. Let row vector $\boldsymbol{\xi}_0 = (\boldsymbol{\xi} : \mathbf{0})_{1\times m}$ be the initial distribution, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{m-k})$, $\sum_{i=1}^{m-k} \xi_i = 1$, and $\mathbf{0} = (0, \ldots, 0)_{1\times k}$.

Theorem 2.2. *For a time-homogeneous Markov chain $\{Y_i\}$ having transition probability matrix of the form in* (2.2.3)*, the probability of the time index t when the chain first enters a set of absorbing states can be obtained from*

$$P(Y_t \in A, Y_{t-1} \notin A, \ldots, Y_1 \notin A \mid \boldsymbol{\xi}_0) = \boldsymbol{\xi}_0 N^{t-1}(I - N)\mathbf{1}' \tag{2.2.4}$$

where $\mathbf{1}_{1\times(m-k)}$ is a unit row vector.

Chapter 3

Single-Server Preemptive Priority Queues Serving One Customer at a Time

In this chapter, we demonstrate the construction of a finite Markov chain to imbed $M/M/1$-$R/b/1$ preemptive priority queues which were meticulously described in Section 1.3 and using the notations introduced in §1.3.2. Let $b < \infty$ be an integer indicating the maximum number of customers, either are in service or are waiting to receive service, allowed to stay in a system at any time. For example, bed-space is very limited in hospital emergency departments. If an ER is over crowded, patients who are not in treatment have to wait outside of ER. If the emergency department is over crowded, then newly arrived patients may have to be deferred to nearby hospitals.

We will show the procedure to follow for obtaining the probability distribution of the variable **wait time** defined slightly different in two ways: in the health care sector, much attention focuses on the time length (1) starting from the time a patient is registered to the time of departure (this time length may also be called the length of stay (LOS); in other queueing systems, the time length (2) starting from the time a customer enters the queue to the time the customer begins receiving service for the first time receive more interest (this time length is called wait time).

To obtain the distribution of the random variable *wait time*, our approach to model a priority queueing system is to monitor information about the queue at discrete time points in increments of Δt and study first the limiting behavior of the system by an imbedded finite Markov chain. For fixed Δt, the long run system behavior can be described by a vector-valued discrete time stochastic process $\{X(t_0 + m\Delta t), m = 0, 1, 2, \cdots\}$ which stores information about the system at the beginning of every interval $[t_0 + m\Delta t, t_0 + (m + 1)\Delta t)$. Without loss of generality and for simplicity, we may let $t_0 = 0$ and write $X(t_0 + m\Delta t) = X_m$. Instead of directly analyzing the continuous time process $X(t)$, we consider a discrete time stochastic process which has vector-valued information $\{X_m = (X^{\mathrm{s}}_m, X^{\mathrm{w}}_{1,m}, \ldots, X^{\mathrm{w}}_{r,m}),\ m = 0, 1, 2, \cdots\}$ where $X^{\mathrm{s}}_m = i$ if a priority i customer is in service at $m\Delta t$, or $X^{\mathrm{s}}_m = 0$ if no one is in service and in the queue. $X^{\mathrm{w}}_{i,m}$ informs the number of priority i customers waiting in queue at $m\Delta t$. Let B_m be the number of customers in the queueing system at time $m\Delta t$ that we will need later when establishing state transition rules and transition probabilities.

3.1 Customers of Urgent and Non-urgent Types

This principle of introducing a discrete priority index parameter to derive the general expected-value formula of machine-repair waiting-line problems is very much applicable to the studying of hospital emergency department wait-line.

For the simplest model of a priority queue in which a customer is categorized either to an urgent or a non-urgent group, let urgent customers be assigned higher service priority over the non-urgent group. Assuming that customers of urgent type arrive to the system at a mean rate λ_1 and non-urgent customers arrive to the system at a mean rate λ_2. The service-time of urgent customers has a mean of $1/\mu_1$ and distribution $F(s_1) = 1 - exp\{-\mu_1 s_1\}$, non-urgent customers has a mean of $1/\mu_2$ and distribution $F(s_2) = 1 - exp\{-\mu_1 s_2\}$. The mean rates of departure from the system is μ_1 and μ_2 for urgent and non-urgent customers, respectively.

Suppose a non-urgent customer is in service and within a time interval of $[t-\Delta t, t)$, the non-urgent customer in service either departs from the queueing system with probability $\mu_2 \Delta t$, or stays in the system with probability $1 - \mu_2 \Delta t$ at time t. If the non-urgent customer stays and, within the same time interval Δt, an urgent customer arrives to the system, then we assume the preemptive-repeat scheduling of the arrived higher priority customer. It means that the service for the non-urgent customer would be interrupted and beginning at time t the arrived urgent customer will take on service. The preempted non-urgent customer would wait at the first position in the non-urgent section of the queue until the departure of any urgent customer in service and there is no urgent customer waiting in line in order to resume to service again. Under such discipline, it is clear that the only time a non-urgent customer can be in service is when there is no urgent customer waiting in line or in service. When the preempted non-urgent customer resumes back to service, the service-time distribution is again exponential with mean $1/\mu_2$. For any other situation regarding the arrival process, service for the non-urgent customer would not be interrupted.

3.1.1 State Space and State Transition Rules

We assume that the queue does not go into a state that no customer is in service while at least one customer is waiting in queue. A finite state space Ω_X is induced by the definition of X_m where Ω_X consists of states satisfying the following set of criteria:

$$\begin{aligned}\Omega_X = \;& (0,0,0) \bigcup \big\{(x^{\mathrm{s}}, x_1^{\mathrm{w}}, x_2^{\mathrm{w}}) \mid x^{\mathrm{s}} = 1 \text{ or } x^{\mathrm{s}} = 2,\ 0 \le x_1^{\mathrm{w}} \le b-1, \\ & \qquad 0 \le x_2^{\mathrm{w}} \le b-1-x_1^{\mathrm{w}},\ x^{\mathrm{s}} = 2 \text{ only if } x_1^{\mathrm{w}} = 0, \text{ and} \\ & \qquad \text{for each } (x^{\mathrm{s}}, x_1^{\mathrm{w}}, x_2^{\mathrm{w}}),\ x^{\mathrm{s}} = 0 \text{ only if both} \\ & \qquad x_1^{\mathrm{w}} = 0 \text{ and } x_2^{\mathrm{w}} = 0\big\}\end{aligned}$$

where

$$x^{\mathrm{s}} = \begin{cases} 0 & \text{if no customer is receiving service,} \\ 1 & \text{if an urgent customer is in service,} \\ 2 & \text{if a non-urgent customer is in service,} \end{cases}$$

x_1^{w} monitors the number of urgent customers waiting in queue and x_2^{w} monitors the number of non-urgent customers waiting in queue.

In help to describe state transitions of the process, we use a bi-variate random variable $(Z_{a,m},\, Z_{d,m})$ to describe the arrival and departure process during $[m\Delta t, (m+1)\Delta t)$ for $m = 0, 1, 2, \cdots$ where

$$Z_{a,m} = \begin{cases} 0 & \text{if no customer arrives to the queue,} \\ 1 & \text{if an urgent subject arrives to the queue,} \\ 2 & \text{if a non-urgent subject arrives to the queue,} \end{cases}$$

$$Z_{d,m} = \begin{cases} 0 & \text{if no customer departs from the system,} \\ x^{\mathrm{s}} & \text{if the customer of service priority } x^{\mathrm{s}} > 0 \\ & \text{in service departs from the system.} \end{cases}$$

For the convenience when describing a one-step transition probability by $p_{\boldsymbol{uv}} = \Pr\{X_{m+1} = \boldsymbol{v} \mid X_m = \boldsymbol{u}\}$ for any m, we use the notation $\boldsymbol{u} \to \boldsymbol{v}$ to describe a one-step state transition of the process going from $X_m = (u^{\mathrm{s}}, u_1^{\mathrm{w}}, u_2^{\mathrm{w}}) = \boldsymbol{u}$ to $X_{m+1} = (v^{\mathrm{s}}, v_1^{\mathrm{w}}, v_2^{\mathrm{w}}) = \boldsymbol{v}$. For a fixed b, we state the transition rules by describing all possible states and their transitions in the following.

When the queue capacity is not reached, $B_m < b$,

$$(0,0,0) \to \quad (Z_{a,m}, 0, 0) \quad \text{if (A1)}$$

$$(1, u_1^{\mathrm{w}}, u_2^{\mathrm{w}}) \to \begin{cases} (1, u_1^{\mathrm{w}}, u_2^{\mathrm{w}}) & \text{if (A2)} \\ (1, u_1^{\mathrm{w}}+1, u_2^{\mathrm{w}}) & \text{if (A3)} \\ (1, u_1^{\mathrm{w}}, u_2^{\mathrm{w}}+1) & \text{if (A4)} \\ (1, u_1^{\mathrm{w}}-1, u_2^{\mathrm{w}}) & \text{if (A5)} \\ (2, 0, u_2^{\mathrm{w}}-1) & \text{if (A6)} \\ (2, 0, u_2^{\mathrm{w}}) & \text{if (A7)} \\ (0,0,0) & \text{if (A8)} \end{cases}$$

$$(2, 0, u_2^{\mathrm{w}}) \to \begin{cases} (1, 0, u_2^{\mathrm{w}}) & \text{if (A9)} \\ (2, 0, u_2^{\mathrm{w}}-1) & \text{if (A10)} \\ (0,0,0) & \text{if (A11)} \\ (2, 0, u_2^{\mathrm{w}}) & \text{if (A12)} \\ (2, 0, u_2^{\mathrm{w}}+1) & \text{if (A13)} \\ (1, 0, u_2^{\mathrm{w}}+1) & \text{if (A14)} \end{cases}$$

where

(A1) the system is empty at time $m\Delta t$ and it stays empty at time $(m+1)\Delta t$ if $Z_{a,m} = 0$, or service begins for any new arrival with service priority score $Z_{a,m} > 0$, note that $Z_{d,m}$ must be zero if X_m is a zero vector,

(A2) $Z_{a,m} = 1$ and $Z_{d,m} = 1$: the urgent customer in service departs from the system and a new urgent customer arrives to the system, or if $Z_{a,m} = 0$ and $Z_{d,m} = 0$: the system stays unchanged since there is no arrival nor departure, (A3) $Z_{a,m} = 1$ and $Z_{d,m} = 0$: the urgent customer in service remains in service, an urgent customer arrives and must wait at the last position in the urgent customer section of the queue,

(A4) $Z_{a,m} = 2$ and $Z_{d,m} = 0$: the urgent customer in service remains in service, a non-urgent customer arrives and must wait at the end of the line for service,

(A5) $Z_{a,m} = 0$ and $Z_{d,m} = 1$ and $u_1^{\mathrm{w}} > 0$: the urgent customer in service departs from the system and there is no new arrival, the urgent customer who waits at the first position begins service next,

(A6) $Z_{a,m} = 0$ and $Z_{d,m} = 1$, $u_1^{\mathrm{w}} = 0$ and $u_2^{\mathrm{w}} > 0$: the urgent customer in service departs from the system, there is no new arrival and no urgent customer waits in line, the non-urgent customer who waits at the first position begins service next,

(A7) $Z_{a,m} = 2$ and $Z_{d,m} = 1$ and $u_1^{\mathrm{w}} = 0$: the urgent customer in service departs from the system, a new non-urgent customer arrives, and there is no urgent customer waiting in line,

(A8) $Z_{a,m} = 0$ and $Z_{d,m} = 1$, $u_1^{\mathrm{w}} = 0$ and $u_2^{\mathrm{w}} = 0$: the urgent customer in service departs from the system, there is no new arrival and no customer waiting in line for service.

(A9) $Z_{a,m} = 1$ and $Z_{d,m} = 2$: the non-urgent customer in service departs from the system and a newly arrived urgent customer begins service next,

(A10) $Z_{a,m} = 0$ and $Z_{d,m} = 2$ and $u_2^{\mathrm{w}} > 0$: the non-urgent customer in service departs from the system and there is no new arrival, the non-urgent customer who waits at the first position in line begins service next,

(A11) $Z_{a,m} = 0$ and $Z_{d,m} = 2$ and $u_2^{\mathrm{w}} = 0$: the non-urgent customer in service departs from the system, there is no new arrival and no customer waiting in line for service,

(A12) $Z_{a,m} = 0$ and $Z_{d,m} = 0$: no arrival and no departure,

(A13) $Z_{a,m} = 2$ and $Z_{d,m} = 0$: the non-urgent customer in service remains in service and a newly arrived non-urgent customer waits at the end of the line for service,

(A14) $Z_{a,m} = 1$ and $Z_{d,m} = 0$: the non-urgent customer in service remains and has service interrupted due to the arrival of an urgent customer, the preempted non-urgent customer waits at the first position in line and service begins for the arrived urgent customer next.

The *preemptive* discipline of the priority queue can be seen in the last state transition rule above. Note that for any m, $u^{\mathrm{s}} = 2$ only if $u_1^{\mathrm{w}} = 0$.

When the system capacity is reached, $B_m = b$, then for any m, $Z_{a,m} = 0$ and

$$(1, u_1^{\mathrm{w}}, u_2^{\mathrm{w}}) \rightarrow \begin{cases} (1, u_1^{\mathrm{w}}, u_2^{\mathrm{w}}) & \text{if (B1)} \\ (1, u_1^{\mathrm{w}} - 1, u_2^{\mathrm{w}}) & \text{if (B2)} \\ (2, 0, u_2^{\mathrm{w}} - 1) & \text{if (B3)} \end{cases}$$

$$(2, 0, u_2^{\mathrm{w}}) \rightarrow \begin{cases} (2, 0, u_2^{\mathrm{w}}) & \text{if (B4)} \\ (2, 0, u_2^{\mathrm{w}} - 1) & \text{if (B5)} \end{cases}$$

where

(B1) $Z_{a,m} = 0$ and $Z_{d,m} = 0$: the system remains unchanged since there is no departure,

(B2) $Z_{a,m} = 0$ and $Z_{d,m} = 1$ and $u_1^{\mathrm{w}} > 0$: the urgent customer in service departs from the system and the urgent customer who waits in the first position begins service next,

(B3) $Z_{a,m} = 0$ and $Z_{d,m} = 1$ and $u_1^{\mathrm{w}} = 0$: the urgent customer in service departs from the system and no urgent customer waits in line, the non-urgent customer who waits in the first position begins service next,

(B4) $Z_{a,m} = 0$ and $Z_{d,m} = 0$: the system remains unchanged since there is no departure,

(B5) $Z_{a,m} = 0$ and $Z_{d,m} = 2$: the non-urgent customer in service departs from the system and there is no urgent customer waiting in line, the non-urgent customer who waits in the first position begins service next.

In our definition, if the system is full at time $m\Delta t$, no new arrival is allowed to enter the system during the time interval $[m\Delta t, (m+1)\Delta t)$ for any m independent of the output process.

3.1.2 Assigning Transition Probabilities and Finding the Ergodic Distribution

For convenience, we let $p_{\boldsymbol{uv}}$ denote the one-step transition probability $\Pr\{X_{m+1} = (v^{\mathrm{s}}, v_1^{\mathrm{w}}, v_2^{\mathrm{w}}) = \boldsymbol{v} \mid X_m = (u^{\mathrm{s}}, u_1^{\mathrm{w}}, u_2^{\mathrm{w}}) = \boldsymbol{u}\}$ for $m = 0, 1, 2, \cdots$. When defining the transition probabilities of the imbedded Markov chain, we need to give heed to the requirement that $\sum_{\boldsymbol{v}} p_{\boldsymbol{uv}} = 1$ for any given $(u^{\mathrm{s}}, u_1^{\mathrm{w}}, u_2^{\mathrm{w}})$.

Assuming that urgent and non-urgent customers arrive to the priority queue as independent Poisson processes with mean rates λ_1 and λ_2, respectively, and assuming that the service-time distributions of both types of customer are independent and are exponentially distributed. The mean rate of departure from the system is μ_1 for urgent customers, and the mean rate is μ_2 for non-urgent customers. For any time m, if there is an arrival of a customer during the time interval $[m\Delta t, (m+1)\Delta t)$, the customer will only be admitted into the queue system if the system is not full at time $m\Delta t$. Suppose there is a customer who, regardless of priority, arrived to the system during the time interval $[m\Delta t, (m+1)\Delta t)$, we do not allow the same customer to depart from the system during $[m\Delta t, (m+1)\Delta t)$ and service may begin for the same customer at a time no sooner than $(m+1)\Delta t$. We list explicitly the state transition probabilities of $M/M/1\text{-}2/b/1$ preemptive repeat-different priority queues in the following.

For any m, if the system is empty, $X_m = (0, 0, 0)$, then

$$p_{\boldsymbol{uv}} = \begin{cases} 1 - \lambda_1 \Delta t - \lambda_2 \Delta t + o(\Delta t) & \text{if } Z_{a,m} = 0, \\ \lambda_{Z_{a,m}} \Delta t + o(\Delta t) & \text{if } Z_{a,m} > 0, \text{ for } Z_{a,m} = 1, 2. \end{cases}$$

If the system is full, $u_m^{\mathrm{s}} > 0$ and $B_m = b$, then

$$p_{\boldsymbol{uv}} = \begin{cases} 1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t) & \text{if } Z_{a,m} = 0 \text{ and } Z_{d,m} = 0, \\ \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t) & \text{if } Z_{a,m} = 0 \text{ and } Z_{d,m} = u^{\mathrm{s}}. \end{cases}$$

If the system is not empty nor full, $u^{\mathrm{s}} > 0$ and $1 \leq B_m < b$, then

$$p_{\boldsymbol{uv}} = \begin{cases} (1 - \lambda_1 \Delta t - \lambda_2 \Delta t + o(\Delta t))\,(1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (1a)} \\ \qquad + \left(\lambda_{Z_{a,m}} \Delta t + o(\Delta t)\right)(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \\ (\lambda_{Z_{a,m}} \Delta t + o(\Delta t))(1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (2a)} \\ & \\ (1 - \lambda_1 \Delta t - \lambda_2 \Delta t + o(\Delta t))(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (3a)} \\ (\lambda_{Z_{a,m}} \Delta t + o(\Delta t))(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (4a)} \end{cases}$$

(1a) $Z_{a,m} = 0$ and $Z_{d,m} = 0$ or $Z_{a,m} = Z_{d,m} = u^{\mathrm{s}}$,

(2a) $Z_{a,m} > 0$ and $Z_{d,m} = 0$, for $Z_{a,m} = 1, 2$,

(3a) $Z_{a,m} = 0$ and $Z_{d,m} = u^{\mathrm{s}}$,

(4a) $Z_{a,m} > 0$ and $Z_{d,m} = u^{\mathrm{s}}$, for $Z_{a,m} = 1, 2$ and $Z_{a,m} \neq u^{\mathrm{s}}$.

3.1.3 Irreducible Chain and Existence of a Unique Ergodic Distribution

To study hospital emergency wait time, we have the assumption that there is a threshold on the maximum number of patients allowed in an emergency department. Therefore, an imbedded Markov chain in describing a hospital emergency department as a priority queueing system always has a finite state space Ω. In practice, it is reasonable to assume that no patient will spend infinite amount of time waiting for and be in a treatment. We further assume that whenever there is available waiting bed-space in an emergency department, new patients continue to be allowed to enter. While the waiting room is temporarily full, newly arrived patients will be re-directed to other hospitals until there is space available again. From every state in Ω, there is a positive probability to reach, in finite amount of time, the *empty state* where all patients received treatment and left the emergency department. From this *empty state*, there is a positive probability that any non-empty state of $X(t) \in \Omega$ can be reached again in a finite amount of time.

In the following, we go through a simple yet a routine argument to show the existence of an ergodic state distribution of our priority queueing model.

For convenience, we write $0 < \lambda_i \Delta t = \lambda_i(\Delta t) < 1$ and $0 < \mu_i \Delta t = \mu_i(\Delta t) < 1$ in this section for the purpose of showing irreducibility of the imbedded finite Markov chains. We may interpret $\lambda_1 = 30$ arrivals per hour and $\Delta t = 1/60$ to express $\lambda_1 \Delta t = 0.5$ arrivals per minute, or simply write $\lambda_1(\Delta t) = 0.5$ arrivals per minute and let Δt be 1 minute intervals as a matter of re-scaling λ_i and μ_i.

We use the vector $(0, 0, 0)$ to denote the empty state when no patient is waiting nor is undergoing treatment, and we assign it to be the initial state to show irreducibility. By the properties of a Poisson arrival process, the one-step transition $(0, 0, 0) \to (0, 0, 0)$ has a transition probability of $p^{[1]}_{(0,0,0)\to(0,0,0)} = (1-\lambda_1(\Delta t)-\lambda_2(\Delta t)+o(\Delta t)) > 0$ and it is not difficult to see that $(0, 0, 0)$ is an aperiodic state. The one-step transition $(0, 0, 0) \to (1, 0, 0)$ has a transition probability of $p^{[1]}_{(0,0,0)\to(1,0,0)} = \lambda_1(\Delta t) + o(\Delta t) > 0$. Similarly the one-step transition probability $p^{[1]}_{(0,0,0)\to(2,0,0)}$ is greater than zero.

For the process to have a two-step transition $(0, 0, 0) \to (1, 0, 0) \to (1, 1, 0)$, by the independent increment property of a Poisson process, the assumption that the arrival process is independent of the departure process and since this is the only path to reach the state $(1, 1, 0)$ in two steps from the empty state, the transition probability must be $p^{[2]}_{(0,0,0)\to(1,1,0)} = \Pr\{(1, 0, 0) \mid (0, 0, 0)\}\times\Pr\{(1, 1, 0) \mid (1, 0, 0)\} > 0$. Similarly, other two-step transition probabilities $p^{[2]}_{(0,0,0)\to(2,0,0)}$, $p^{[2]}_{(0,0,0)\to(1,1,0)}$, $p^{[2]}_{(0,0,0)\to(1,0,1)}$, and $p^{[2]}_{(0,0,0)\to(2,0,1)}$ all can easily be shown to be greater than zero.

Furthermore, suppose we are interested in the three-step transition of entering the state $(1, 1, 1)$, the state when an urgent patient is undergoing treatment and an urgent and a non-urgent patient are waiting in line, one possible path to reach this state from the empty state is $(0, 0, 0) \to (1, 0, 0) \to (1, 1, 0) \to (1, 1, 1)$ with a probability $\Pr\{(1, 0, 0) \mid (0, 0, 0)\}\times\Pr\{(1, 1, 0) \mid (1, 0, 0)\}\times\Pr\{(1, 1, 1) \mid (1, 1, 0)\} > 0$. However, since there are other possible paths to reach $(1, 1, 1)$ in three steps from the empty state, ie. $(0,0,0) \to (2,0,0) \to (1,0,1) \to (1,1,1)$, therefore we have the inequality $p^{[3]}_{(0,0,0)\to(1,1,1)} \geq \Pr\{(1,0,0) \mid (0,0,0)\} \times \Pr\{(1,1,0) \mid (1,0,0)\} \times \Pr\{(1,1,1) \mid (1,1,0)\} > 0$ which is a direct result from the Chapman-Kolmogorov identity.

Similarly, for any nonempty state $(1, x_1^{\mathrm{w}}, x_2^{\mathrm{w}})$ or $(2, 0, x_2^{\mathrm{w}})$, and under the conditions $(x_1^{\mathrm{w}} + x_2^{\mathrm{w}}) \leq (b-1)$, $x_1^{\mathrm{w}} \geq 0$ and $x_2^{\mathrm{w}} \geq 0$, it is not difficult to see that there exists an integer m, $(x_1^{\mathrm{w}} + x_2^{\mathrm{w}}) \leq m < \infty$, that in m steps, $p^{[m]}_{(0,0,0)\to(1,x_1^{\mathrm{w}},x_2^{\mathrm{w}})} > 0$ or $p^{[m]}_{(0,0,0)\to(2,0,x_2^{\mathrm{w}})} > 0$.

It is trivial that the state $(0,0,0)$ is aperiodic and all other non-empty states communicate with it. All states can reach one another in a finite number of steps with positive probabilities. By the following criterion and theorem:

Criterion 3.0.1. *A chain is irreducible if, and only if, every state can be reached from every other state.*

Theorem 3.1. *In an irreducible Markov chain, all states belong to the same class: they are all transient, all persistent null states, or all persistent non-null states. In every case they have the same period. Moreover, every state can be reached from every other state.* $\cdots$

as stated in Feller [9] Chapter XV Section 4 and Section 5, all states in the finite state space Ω_X are therefore aperiodic and persistent. It is intuitive that there exists no null or transient states in a finite irreducible aperiodic Markov chain. By the Theorem (b) in Chapter XV section 6, Feller [9] also showed that there exists an unique stationary distribution with no zeros (ergodic distribution) for the states in

Ω_X. Alternatively, results of Isaacson and Madsen [17] can also be used to show the existence of a unique ergodic distribution (or steady-state) distribution.

3.1.4 Obtaining the Unique Ergodic Distribution

We show in this section that there are two methods in obtaining the ergodic distribution given a transition probability matrix $M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})$, $k < \infty$, in our setting.

First Algorithm

Lemma 3.1.1. *If all entries of a matrix A are positive, $a_{ij} > 0$, and $\boldsymbol{v}'$ is an eigenvector of A with $v_j \geq 0$ for all j, then all entries of $\boldsymbol{v}'$ are strictly positive.*

Proof for the above Lemma is trivial.

Lemma 3.1.2. *If a matrix A with all entries non-negative, but A is not a zero matrix, has a right eigenvector $\boldsymbol{v}'$ (a column vector), having only positive elements, associated with eigenvalue λ, then every left eigenvector $\boldsymbol{u}$ (a row vector) of A, having only non-negative elements, also has the same eigenvalue λ.*

Proof. Since $\boldsymbol{u}$ and $\boldsymbol{v}'$ can not be zero vectors and $\boldsymbol{v}'$ has only positive elements and some elements of $\boldsymbol{u}$ are positive, we note that $\boldsymbol{u}\boldsymbol{v}' > 0$. Suppose $\boldsymbol{u}$ has eigenvalue λ^*, we have

$$\lambda \boldsymbol{u}\boldsymbol{v}' = \boldsymbol{u}(A\boldsymbol{v}') = (\boldsymbol{u}A)\boldsymbol{v}' = \lambda^* \boldsymbol{u}\boldsymbol{v}'.$$

So the eigenvalues λ^* must be equal to λ. □

Theorem 3.2. *(Perron-Frobenius) Let $A = [a_{ij}]$ be a real $n \times n$ matrix and is non-negative and irreducible, then the following statements hold:*

1. *there is a positive real eigenvalue λ of A such that any other eigenvalue λ' of A satisfy $\mid \lambda' \mid < \lambda$;*

2. *there is a left (or respectively a right) eigenvector, associated with λ of A, having all positive elements;*

3. *λ is simple, or has algebraic multiplicity 1.*

For the proof of the above theorem, please see Theorem 1 in Lancaster and Tismenetsky [23] Chapter 15 Section 3.

One property of a transition probability matrix $M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is that all row sums are one, so we have

$$M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) \cdot \mathbf{1}' = \mathbf{1}'$$

where $\mathbf{1}$ is a $k\times 1$ unit row vector. By Lemma 3.1.2, every left non-negative eigenvector of $M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})$ has eigenvalue 1. By Theorem 3.2, we see that the absolute values of all other eigenvalues of $M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is less than $\lambda = 1$ in this case. Since λ has algebraic multiplicity 1, there exist only one normalized left eigenvector $\boldsymbol{u}_{normal}$ of the transition probability matrix $M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})$ associated with eigenvalue 1 that is the ergodic distribution having only positive elements.

Second Algorithm

Under closer examination, transition probability matrices defining priority queues under the Poisson arrival and exponential service assumption in this thesis have a particular form $M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) = I_{k\times k} + \Delta t \cdot M_{k\times k}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ if numerically the function $o(\Delta t)$ wherever it appears is replaced by 0. $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ can be singular or vectors of positive parameters, $I_{k\times k}$ is an identity matrix and $M_{k\times k}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is a non-zero matrix.

Theorem 3.3. *Given that a finite irreducible aperiodic Markov chain having transition probability matrix $M_{k\times k}(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) = I_{k\times k} + \Delta t \cdot M_{k\times k}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and there exists an ergodic distribution, such distribution is independent of Δt.*

Proof. Suppose we have two transition probability matrices $M(\Delta t,\ \boldsymbol{\lambda},\ \boldsymbol{\mu})$ and $M(\Delta\tau,\ \boldsymbol{\lambda},\ \boldsymbol{\mu})$ that differ only in Δt and $\Delta\tau$, $\Delta t > 0$, $\Delta\tau > 0$ and $\Delta t \neq \Delta\tau$. Suppose

$$\begin{aligned}
\boldsymbol{u}M(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \boldsymbol{u} \\
\boldsymbol{v}M(\Delta\tau, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \boldsymbol{v} \\
\boldsymbol{u}\cdot\mathbf{1}' &= 1 \\
\boldsymbol{v}\cdot\mathbf{1}' &= 1 \\
\boldsymbol{u} &= (u_1, u_2, \cdots) \\
\boldsymbol{v} &= (v_1, v_2, \cdots) \\
\mathbf{1} &= (1, 1, \cdots)
\end{aligned} \tag{3.1.1}$$

$\boldsymbol{u}$ and $\boldsymbol{v}$ are non-negative vectors and $\mathbf{1}$ is a unit row vector of proper length.

Since

$$\boldsymbol{u} \cdot M^m(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{u}$$

for $m = 1, 2, \cdots$ trivially, then

$$\boldsymbol{u} \cdot \lim_{m\to\infty} M^m(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{u}$$

where $\boldsymbol{u}$ is an ergodic distribution of the finite Markov chain and $\lim\limits_{m\to\infty} M^m(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is of identical rows of $\boldsymbol{u}$. $\boldsymbol{u}$ is unique, see Feller [9]. Same can we say about $\boldsymbol{v}$ with respect to $\lim\limits_{m\to\infty} M^m(\Delta\tau, \boldsymbol{\lambda}, \boldsymbol{\mu})$. It suffices to show that $\boldsymbol{u} = \boldsymbol{v}$.

From (3.1.1), we have

$$\begin{aligned} \boldsymbol{u} =& \ \boldsymbol{u} \cdot M(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{u}\ (I + \Delta t M(\boldsymbol{\lambda}, \boldsymbol{\mu})) \\ \boldsymbol{u} =& \ \boldsymbol{u} + \boldsymbol{u} \cdot \Delta t \cdot M(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \boldsymbol{u} \cdot M(\boldsymbol{\lambda}, \boldsymbol{\mu}) =& \ \mathbf{0} \end{aligned}$$

By the same token, we can show that $\boldsymbol{v}^T\ M(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}^T$. Thus,

$$\begin{aligned} \boldsymbol{v} \cdot M(\boldsymbol{\lambda}, \boldsymbol{\mu}) =& \ \mathbf{0} \\ \boldsymbol{v} \cdot \Delta t M(\boldsymbol{\lambda}, \boldsymbol{\mu}) =& \ \mathbf{0} \\ \boldsymbol{v} + \boldsymbol{v} \cdot \Delta t \cdot M(\boldsymbol{\lambda}, \boldsymbol{\mu}) =& \ \boldsymbol{v} \\ \boldsymbol{v} \cdot (I + \Delta t M(\boldsymbol{\lambda}, \boldsymbol{\mu})) =& \ \boldsymbol{v} \cdot M(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{v} \end{aligned}$$

and trivially

$$\boldsymbol{v} \cdot \lim_{m\to\infty} M^m(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{v}.$$

$\boldsymbol{v}$ is an ergodic distribution of the same finite irreducible aperiodic Markov chain, and by the uniqueness of $\boldsymbol{u}$ with respect to $\lim\limits_{m\to\infty} M^m(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})$, we have $\boldsymbol{v} = \boldsymbol{u}$ and our theorem is proved.

Furthermore, it is easy to see that $\operatorname{rank}(M(\Delta t, \boldsymbol{\lambda}, \boldsymbol{\mu})) = \operatorname{rank}(M(\boldsymbol{\lambda}, \boldsymbol{\mu})) = k - 1$ and $\boldsymbol{u}$ can be determined by $M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ in solving the following system of equations:

$$\begin{aligned} \boldsymbol{u} \cdot M(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \mathbf{0} \\ \boldsymbol{u} \cdot \mathbf{1}' &= 1 \end{aligned} \tag{3.1.2}$$

From (3.1.2), we get

$$\boldsymbol{u} \cdot M^*(\boldsymbol{\lambda}, \boldsymbol{\mu}) = (0, \dots, 0, 1)$$

where, for convenience, $M^*(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is $M(\boldsymbol{\lambda}, \boldsymbol{\mu})$ with the last column added by a $k \times 1$ unit vector. Now $M^*(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is of full rank and

$$\boldsymbol{u} = (0, \dots, 0, 1)(M^*(\boldsymbol{\lambda}, \boldsymbol{\mu}))^{-1}$$

can be obtained to be the last row of the inverse of $M^*(\boldsymbol{\lambda}, \boldsymbol{\mu})$. □

Example 3.1.1. To illustrate the method to obtain the ergodic distribution of a two class preemptive priority queue which may be used to model ER service, we synthesize the parameter values in this numerical example. Suppose that arrived patients are to be categorized either as a priority I (urgent) or a priority II (less urgent) patient. The hospital does not allow more than three patients to occupy the emergency department, either in treatment or are waiting for treatment, and only one patient at a time can be receiving treatment. In our notation, we are modeling a $M/M/1$-$2/3/1$ priority queue. Furthermore, suppose priority I and II patients arrive to the hospital at mean rates of $\lambda_1 = 1$ and $\lambda_2 = 3$ per two hours, respectively. The mean treatment times for priority II and II patients are $\mu_1 = 80$, $\mu_2 = 30$ minutes.

Let Δt be 10-minute intervals. The state space consists of the states

$$\{(0,0,0),(1,0,0),(1,1,0),(1,2,0),(1,0,1),(1,1,1),(1,0,2),(2,0,0),(2,0,1),(2,0,2)\}.$$

By Section 3.1.2, the transition probability matrix corresponding to the above finite state space can be constructed to be

$$\begin{array}{c|cccccccccc}
(0,0,0) & 0.667 & 0.083 & 0 & 0 & 0 & 0 & 0 & 0.250 & 0 & 0 \\
(1,0,0) & 0.083 & 0.594 & 0.073 & 0 & 0.219 & 0 & 0 & 0.031 & 0 & 0 \\
(1,1,0) & 0 & 0.083 & 0.594 & 0.073 & 0.031 & 0.219 & 0 & 0 & 0 & 0 \\
(1,2,0) & 0 & 0 & 0.125 & 0.875 & 0 & 0 & 0 & 0 & 0 & 0 \\
(1,0,1) & 0 & 0 & 0 & 0 & 0.594 & 0.073 & 0.219 & 0.083 & 0.031 & 0 \\
(1,1,1) & 0 & 0 & 0 & 0 & 0.125 & 0.875 & 0 & 0 & 0 & 0 \\
(1,0,2) & 0 & 0 & 0 & 0 & 0 & 0 & 0.875 & 0 & 0.125 & 0 \\
(2,0,0) & 0.222 & 0 & 0 & 0 & 0.083 & 0 & 0 & 0.528 & 0.167 & 0 \\
(2,0,1) & 0 & 0 & 0 & 0 & 0 & 0 & 0.083 & 0.222 & 0.528 & 0.167 \\
(2,0,2) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.333 & 0.667
\end{array}$$

For limited spacing, the transition probabilities were reported to the nearest third decimal. The ergodic state distribution $\boldsymbol{\pi}$ can be obtained to be

$$\boldsymbol{\pi} = (0.1225, 0.0263, 0.0058, 0.0034, 0.0651, 0.0480, 0.2496, 0.1739, 0.2036, 0.1018).$$

□

3.2 More Than Two Priority Classes

After seeing examples of how transition probability matrix and the steady state distributions can be constructed for a two class priority queueing model, we now extend our work to preemptive-repeat priority queues that have customers of more than two categories. Using our notation, we study $M/M/1$-$R/b/1$ preemptive priority queues. Suppose that each customer arriving to the system can first be classified into one of $1 < R < \infty$ priority classes. For fixed b, the finite state space Ω_X consists of states which satisfy the following criteria:

$$\begin{aligned}\Omega_X = \quad & \{\mathbf{0}\} \bigcup \big\{X = (x^{\mathrm{s}}, x_1^{\mathrm{w}}, \ldots, x_R^{\mathrm{w}}) \mid x^{\mathrm{s}} \in \{1, \ldots, R\}, \\ & \qquad 0 \le x_1^{\mathrm{w}} \le b-1,\ 0 \le x_i^{\mathrm{w}} \le b - 1 - \sum_{j=1}^{i-1} x_j^{\mathrm{w}} \\ & \qquad \text{for } i = 2, \ldots, R, \text{ and } x_i^{\mathrm{w}} = 0 \text{ for all} \\ & \qquad i = 1, \ldots, x^{\mathrm{s}} - 1\big\}\end{aligned}$$

where $\mathbf{0}$ is a zero vector of proper length, x^{s} monitors the priority class of the customer in service, x^{s} is in $\{1, \ldots, R\}$ or $x^{\mathrm{s}} = 0$ if no one is in service, and x_i^{w} monitors the number of priority i customers waiting in queue, $i = 1, 2, \ldots, R$.

Let $(Z_{a,m}, Z_{d,m})$ be defined the same way as in the previous. $Z_{a,m}$ is in $\{0, \ldots, R\}$ and $Z_{d,m}$ can either be X_m^{s} or be 0 during $[m\Delta t, (m+1)\Delta t)$ for all m. Assuming that priority class i customers arrive to the system as a Poisson process with rate λ_i and is independent of all other classes for $i = 1, 2, \ldots, R$. Under the same preemptive-repeat priority discipline and the assumptions in Section §3.1, we may generalize the state transition rules and probabilities, for $R > 2$, as in the following for the purpose of obtaining the ergodic state distribution.

3.2.1 Transition Rules and Transition Probabilities for Preemptive Priority Queues Allowing up to One Customer in Service

For the convenience of describing one-step state transitions, we let $(X^{\mathrm{s}}_m, X^{\mathrm{w}}_{1,m}, \ldots, X^{\mathrm{w}}_{R,m}) = (u^{\mathrm{s}}, u^{\mathrm{w}}_1, \ldots, u^{\mathrm{w}}_R) = \boldsymbol{u}$ and $(X^{\mathrm{s}}_m, X^{\mathrm{w}}_{1,m+1}, \ldots, X^{\mathrm{w}}_{R,m+1}) = (v^{\mathrm{s}}, v^{\mathrm{w}}_1, \ldots, v^{\mathrm{w}}_R) = \boldsymbol{v}$. In this chapter, let $K^*_m = \min\{k : X^{\mathrm{w}}_{k,m} > 0, k = 1, \ldots, R\}$. K^*_m essentially is the priority class of the customer waiting at the first position in line at time $m\Delta t$.

We establish the state transition rules using a more compact form to express state transitions. For fixed b and any $m \in \mathbb{N}_0$, if the system is completely empty, $(X^{\mathrm{s}}_m, X^{\mathrm{w}}_{1,m}, \ldots, X^{\mathrm{w}}_{R,m}) = \mathbf{0}$, then $Z_{d,m} = 0$ and

$$\begin{aligned} v^{\mathrm{s}} = &\ I_{Z_{a,m}>0}(Z_{a,m}) \times Z_{a,m} \\ v^{\mathrm{w}}_i = &\ 0 \text{ for } i = 1, \ldots, R. \end{aligned}$$

If the system has a customer in service such that there may or may not be customers waiting in line and the system is not full, $u^{\mathrm{s}} > 0$ and $0 \leq \sum_{i=1}^{R} u^{\mathrm{w}}_i < b - 1$, then

$$v^{\mathrm{s}} = \begin{cases} u^{\mathrm{s}} & \text{if } Z_{d,m} = 0 \text{ and } Z_{d,m} = 0, \\ \min(Z_{a,m}, u^{\mathrm{s}}) & \text{if } Z_{d,m} = 0 \text{ and } Z_{a,m} > 0, \\ Z_{a,m} & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and } \sum_{i=1}^{R} u_i^{\mathrm{w}} = 0, \\ K_m^* & \text{if } Z_{d,m} = u^{\mathrm{s}},\ Z_{a,m} = 0 \text{ and } \sum_{i=1}^{R} u_i^{\mathrm{w}} > 0, \\ \min(Z_{a,m}, K_m^*) & \text{if } Z_{d,m} = u^{\mathrm{s}},\ Z_{a,m} > 0 \text{ and } \sum_{i=1}^{R} u_i^{\mathrm{w}} > 0 \end{cases}$$

The transition rule of the ℓth entry in $(u_1^{\mathrm{w}}, \ldots, u_R^{\mathrm{w}})$ for $\ell = 1, \ldots, R$ can be expressed as

$$v_\ell^{\mathrm{w}} = u_\ell^{\mathrm{w}} + I_1(Z_{a,m}, Z_{d,m}, \ell) - I_2(Z_{a,m}, Z_{d,m}, \ell)$$

where

$$I_1(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} = 0,\ Z_{a,m} > 0 \text{ and } \ell = \max(u^{\mathrm{s}}, Z_{a,m}), \text{ or} \\ & \text{if } Z_{d,m} = u^{\mathrm{s}},\ \sum_{i=1}^{R} u_i^{\mathrm{w}} > 0,\ Z_{a,m} \geq K_m^* \text{ and } \ell = Z_{a,m}, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$I_2(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} = u^{\mathrm{s}},\ \sum_{i=1}^{R} u_i^{\mathrm{w}} > 0,\ Z_{a,m} = 0 \text{ or } Z_{a,m} \geq K_m^*, \\ & \text{and } \ell = K_m^*, \\ 0 & \text{otherwise,} \end{cases}$$

are indicator functions to control whether the number of customer of the ith priority group has been increased, decreased by a unit, or stays unchanged from $m\Delta t$ to $(m+1)\Delta t$.

Else, if the system capacity is reached at time $m\Delta t$, meaning $u^{\mathrm{s}} > 0$ and $\sum_{i=1}^{R} u_i^{\mathrm{w}} = b - 1$, then for sure $Z_{a,m} = 0$ and

$$v^{\mathrm{s}} = \begin{cases} u^{\mathrm{s}} & \text{if } Z_{d,m} = 0, \\ K_m^* & \text{if } Z_{d,m} = u^{\mathrm{s}}, \end{cases}$$

$$v_\ell^{\mathrm{w}} = u_\ell^{\mathrm{w}} - I(Z_{d,m}, \ell)$$

where

$$I(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and } \ell = K_m^*, \\ 0 & \text{otherwise.} \end{cases}$$

For any m and fixed value of b, given $X_m = \boldsymbol{u}$, the one-step state transition probabilities of the imbedded Markov chain can be generalized as in the following.

If $(X_m^{\mathrm{s}}, X_{1,m}^{\mathrm{w}}, \ldots, X_{R,m}^{\mathrm{w}}) = \mathbf{0}$, then

$$p_{\boldsymbol{uv}} = \begin{cases} 1 - \sum_{i=1}^{R} \lambda_i \Delta t + o(\Delta t) & \text{if } Z_{a,m} = 0, \\ \lambda_{Z_{a,m}} \Delta t + o(\Delta t) & \text{if } Z_{a,m} > 0 \text{ for } Z_{a,m} = 1, \ldots, R. \end{cases}$$

If $u^{\mathrm{s}} > 0$ and $\sum_{i=1}^{R} u_i^{\mathrm{w}} = b - 1$, then

$$p_{\boldsymbol{uv}} = \begin{cases} 1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t) & \text{if } Z_{a,m} = 0 \text{ and } Z_{d,m} = 0, \\ \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t) & \text{if } Z_{d,m} = u^{\mathrm{s}}. \end{cases}$$

If $u^{\mathrm{s}} > 0$ and $0 \leq \sum_{i=1}^{R} u_i^{\mathrm{w}} < b-1$, then

$$p_{\boldsymbol{uv}} = \begin{cases} \left(1 - \sum_{i=1}^{R} \lambda_i \Delta t + o(\Delta t)\right)(1 - \mu_{u^{\mathrm{s}}}\Delta t + o(\Delta t)) & \text{if (1b)} \\ \quad +(\lambda_{Z_{a,m}}\Delta t + o(\Delta t))(\mu_{u^{\mathrm{s}}}\Delta t + o(\Delta t)) & \\ (\lambda_{Z_{a,m}}\Delta t + o(\Delta t))(1 - \mu_{u^{\mathrm{s}}}\Delta t + o(\Delta t)) & \text{if (2b)} \\ \left(1 - \sum_{i=1}^{R} \lambda_i \Delta t + o(\Delta t)\right)(\mu_{u^{\mathrm{s}}}\Delta t + o(\Delta t)) & \text{if (3b)} \\ (\lambda_{Z_{a,m}}\Delta t + o(\Delta t))(\mu_{u^{\mathrm{s}}}\Delta t + o(\Delta t)) & \text{if (4b)} \end{cases}$$

where

(1b) $Z_{a,m} = 0$ and $Z_{d,m} = 0$, or $Z_{a,m} = u^{\mathrm{s}}$ and $Z_{d,m} = u^{\mathrm{s}}$,

(2b) $Z_{a,m} > 0$ and $Z_{d,m} = 0$, for $Z_{a,m} = 1, \ldots, R$,

(3b) $Z_{a,m} = 0$ and $Z_{d,m} = u^{\mathrm{s}}$,

(4b) $Z_{a,m} > 0$ and $Z_{d,m} = u^{\mathrm{s}}$, for $Z_{a,m} = 1, \ldots, R$ and $Z_{a,m} \neq u^{\mathrm{s}}$.

Once the transition probability matrix is obtained, the ergodic distribution of the system can easily be calculated as we have shown for the two-priority preemptive-repeat queues. In the next section, we will show the imbedding procedure to obtain the tail distribution of the random variable, length of stay, which is commonly recorded in databases of emergency medical centres.

3.2.2 Finding the Distribution of the Variable *Length of Stay* (LOS) from the Completion of Triage and Registration to the Time of Patient Departure

In the application of hospital emergency patient treatment, suppose our focus is on the probability distribution of the total duration of stay of a particular priority r $(1 \leq r \leq R)$ patient who arrives to the hospital during $[t_0 - \Delta t, t_0)$ for some t_0. In other words, we want to obtain the tail distribution of the time length from t_0 to the time of departure of the same priority r patient, say t_α, from the emergency hospital. Using the FMCI technique, we need to re-imbed a Markov chain starting at time t_0. Redefining the state transition rules and transition probabilities is then necessary in order to obtain the tail distribution of the variable LOS.

With the same queueing and service discipline assumed as described in Sections 3.1 and 3.2, now we use a new vector $(Y_m^{\mathrm{s}}, Y_{1,m}^{\mathrm{w}}, \ldots, Y_{r,m}^{\mathrm{w}}, Y_{r+1,m}^{\mathrm{w}})$ to store information of the queue system for a fixed r for $m = 0, 1, \cdots$. Since our attention on the system begins at time t_0 and at least there should be one patient with priority level r in the system, for sure we know $Y_0^{\mathrm{s}} > 0$. For $m > 0$, entries of the vector $(Y_m^{\mathrm{s}}, Y_{1,m}^{\mathrm{w}}, \ldots, Y_{r,m}^{\mathrm{w}}, Y_{r+1,m}^{\mathrm{w}})$ monitors the following information: $Y_m^{\mathrm{s}} \in \{1, \ldots, r\}$ monitors the priority score of the patient in treatment at time $t_0 + m\Delta t$; $Y_{i,m}^{\mathrm{w}}$ monitors the number of priority i, for $i = 1, 2, \ldots, r$, patients waiting to receive treatment at time $t_0 + m\Delta t$; and $Y_{r+1,m}^{\mathrm{w}}$ monitors: (1) the total number of patients who have priority scores larger than r and are waiting for treatment when $m = 0$; and (2) the total number of patients with priority scores larger than r waiting for treatment at time $t_0 + m\Delta t$ for $m = 1, 2, \cdots$ plus the total number of priority r patients admitted after time t_0. Note the difference in information which $Y^{\mathrm{w}}_{r+1,0}$ and $Y^{\mathrm{w}}_{r+1,m}$ stores for $m = 1, 2, \cdots$.

By the definition of the process Y_{rm} for $m = 0, 1, \cdots$, a finite state space Ω_r also is induced consisting of $\boldsymbol{\alpha}_r$ and all states which satisfy the following criteria:

$$\begin{aligned}\Omega_r = \quad & \{\boldsymbol{\alpha}_r\} \bigcup \{(y^{\mathrm{s}}, y^{\mathrm{w}}_1, \ldots, y^{\mathrm{w}}_{r+1}) \mid 1 \le y^{\mathrm{s}} \le r,\ 0 \le y^{\mathrm{w}}_1 \le b-1, \\ & \qquad 0 \le y^{\mathrm{w}}_2 \le b-1-y^{\mathrm{w}}_1,\ 0 \le y^{\mathrm{w}}_i \le b-1-\sum_{j=1}^{i-1} y^{\mathrm{w}}_j \\ & \qquad \text{for } i = 3, \ldots, r+1 \text{ and for each } (y^{\mathrm{s}}, y^{\mathrm{w}}_1, \ldots, y^{\mathrm{w}}_{r+1}), \\ & \qquad y^{\mathrm{w}}_j = 0 \text{ for all } j = 1, \ldots, y^{\mathrm{s}}-1 \text{ when } y^{\mathrm{s}} > 1\}\end{aligned}$$

where $\boldsymbol{\alpha}_r$ is a vector state denoting the absorbing state of the process. It is a state used to indicate the departure of the priority r patient of interest from the emergency department.

Note that for any m, $y^{\mathrm{s}} = 0$ only if $y^{\mathrm{w}}_i = 0$ for all $i = 1, \ldots, r$. When $\{Y_{rm}, m = 1, 2, \cdots\}$, has reached $\mathbf{0}$ for the first time, the system has entered the absorbing state $\boldsymbol{\alpha}_r$. Our main problem can be properly formulated as to obtaining the tail probability distribution of the variable $W(\boldsymbol{\alpha}_r)$, such that the absorbing state is reached for the first time after t_0. Observe that for fixed R, b, and r, $Y^{\mathrm{w}}_{r+1,m}$ is non-decreasing from time t_0 and subsequent times $t_0 + m\Delta t$ for $m = 1, 2, \cdots$. Furthermore, let the arrival-departure information $(Z_{a,m}, Z_{d,m})$ be defined as before, and for the convenience in describing one step transition rules and state transition probabilities of the stochastic process describing the queue, we suppose given $(Z_{a,m}, Z_{d,m})$ and $(Y^{\mathrm{s}}_{m}, Y^{\mathrm{w}}_{1,m}, \ldots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m}) = (u^{\mathrm{s}}, u^{\mathrm{w}}_{1}, \ldots, u^{\mathrm{w}}_{r}, u^{\mathrm{w}}_{r+1})$ the process will make a transition to a state $(Y^{\mathrm{s}}_{m+1}, Y^{\mathrm{w}}_{1,m+1}, \ldots, Y^{\mathrm{w}}_{r,m+1}, Y^{\mathrm{w}}_{r+1,m+1}) = (v^{\mathrm{s}}, v^{\mathrm{w}}_{1}, \ldots, v^{\mathrm{w}}_{r}, v^{\mathrm{w}}_{r+1})$ for any $m \in \mathbb{N}_0$.

Transition Rules for Obtaining the Waiting-Time Distribution of LOS

When the system has only the particular priority r patient of interest undergoing treatment and there is no patient waiting in line with priority score less than r nor is the threshold on the maximum number of patients allowed in the system reached, that is, $(Y^{\mathrm{s}}_{m}, Y^{\mathrm{w}}_{1,m}, \ldots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m}) = (u^{\mathrm{s}}, 0, \ldots, 0, u^{\mathrm{w}}_{r+1})$ where $u^{\mathrm{s}} = r$ and $u^{\mathrm{w}}_{r+1} < b - 1$, then

$$v^{\mathrm{s}} = \begin{cases} r & \text{if } Z_{a,m} = 0 \text{ and } Z_{d,m} = 0, \\ \min(Z_{a,m}, r) & \text{if } Z_{a,m} > 0 \text{ and } Z_{d,m} = 0, \\ \alpha^{\mathrm{s}} & \text{if } Z_{d,m} = r, \end{cases}$$

for $\ell = 1, \ldots, r-1$,

$$v^{\mathrm{w}}_{\ell} = \begin{cases} 0 & \text{for any } Z_{a,m} \text{ and } Z_{d,m} = 0, \\ \alpha^{\mathrm{w}}_{\ell} & \text{if } Z_{d,m} = r, \end{cases}$$

$$v^{\mathrm{w}}_{r} = \begin{cases} 0 & \text{if } Z_{d,m} = 0 \text{ and, } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ 1 & \text{if } Z_{d,m} = 0 \text{ and } 0 < Z_{a,m} < r, \\ \alpha^{\mathrm{w}}_{r} & \text{if } Z_{d,m} = r, \end{cases}$$

$$v^{\mathrm{w}}_{r+1} = \begin{cases} u^{\mathrm{w}}_{r+1} + 1 & \text{if } Z_{d,m} = 0 \text{ and } Z_{a,m} \geq r, \\ u^{\mathrm{w}}_{r+1} & \text{if } Z_{d,m} = 0 \text{ and } Z_{a,m} < r, \\ \alpha^{\mathrm{w}}_{r+1} & \text{if } Z_{d,m} = r, \end{cases}$$

where α^{s} and $\alpha^{\mathrm{w}}_{\ell}$ for $\ell = 1, \ldots, r+1$ are the components of the absorbing state $\boldsymbol{\alpha}_r = (\alpha^{\mathrm{s}}, \alpha^{\mathrm{w}}_1, \ldots, \alpha^{\mathrm{w}}_r, \alpha^{\mathrm{w}}_{r+1})$.

If the particular priority r patient of interest has to wait in line after arriving to the system, and the threshold on the maximum number of patients allowed in the system is not reached, that is, $0 < u^{\mathrm{s}} \leq r$, $u^{\mathrm{w}}_r > 0$ and $\sum\limits_{i=1}^{r+1} u^{\mathrm{w}}_i < b - 1$, then for $m = 1, 2, \cdots$,

$$v^{\mathrm{s}} = \begin{cases} u^{\mathrm{s}} & \text{if } Z_{d,m} = 0, \text{ and } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ \min(Z_{a,m}, u^{\mathrm{s}}) & \text{if } Z_{d,m} = 0 \text{ and } 0 < Z_{a,m} < r, \\ K^*_m & \text{if } Z_{d,m} = y^{\mathrm{s}} \text{ and } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ \min(Z_{a,m}, K^*_m) & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and } 0 < Z_{a,m} < r. \end{cases}$$

For $\ell = 1, \ldots, r-1$,

$$v^{\mathrm{w}}_{\ell} = u^{\mathrm{w}}_{\ell} + I_1(Z_{d,m}, Z_{a,m}, \ell) - I_2(Z_{d,m}, Z_{a,m}, \ell)$$

where

$$I_1(Z_{d,m}, Z_{a,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} = 0,\ 0 < Z_{a,m} < r,\ u^{\mathrm{s}} < r \text{ and} \\ & \ell = \max(u^{\mathrm{s}}, Z_{a,m}), \text{ or} \\ & \text{if } Z_{d,m} = u^{\mathrm{s}},\ 0 < Z_{a,m} < r \text{ and } \ell = Z_{a,m}, \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{d,m}, Z_{a,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} = u^{\text{s}},\ Z_{a,m} = 0 \text{ and } \ell = K^*_m, \text{ or} \\ & \text{if } Z_{d,m} = u^{\text{s}},\ 0 < Z_{a,m} < r, \text{ and } \ell = \min(K^*_m, Z_{a,m}), \\ 0 & \text{otherwise,} \end{cases}$$

and

$$v^{\text{w}}_r = \begin{cases} u^{\text{w}}_r - 1 & \text{if } Z_{d,m} = u^{\text{s}},\ \sum\limits_{i=1}^{r-1} u^{\text{w}}_i = 0, \text{ and } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ u^{\text{w}}_r + 1 & \text{if } u^{\text{s}} = r,\ Z_{d,m} = 0, \text{ and } 0 < Z_{a,m} < r, \\ u^{\text{w}}_r & \text{otherwise,} \end{cases}$$

$$v^{\text{w}}_{r+1} = \begin{cases} u^{\text{w}}_{r+1} + 1 & \text{if } Z_{a,m} \geq r, \\ u^{\text{w}}_{r+1} & \text{otherwise.} \end{cases}$$

If the threshold on the maximum number of patients allowed in the system is reached, $u^{\text{w}}_r > 0$ and $\sum\limits_{i=1}^{r+1} u^{\text{w}}_i = b - 1$, then

$$v^{\text{s}} = \begin{cases} u^{\text{s}} & \text{if } Z_{d,m} = 0, \\ K^*_m & \text{if } Z_{d,m} = u^{\text{s}}, \end{cases}$$

for $\ell = 1, \ldots, r$,

$$v^{\text{w}}_\ell = u^{\text{w}}_\ell - I(Z_{d,m}, \ell)$$

where

$$I(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} = u^{\text{s}} \text{ and } \ell = K^*_m \\ 0 & \text{otherwise,} \end{cases}$$

and

$$v^{\text{w}}_{r+1} = u^{\text{w}}_{r+1}.$$

If $(Y^{\mathrm{s}}_{m}, Y^{\mathrm{w}}_{1,m}, \dots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m}) = (u^{\mathrm{s}}, 0, \dots, 0, u^{\mathrm{w}}_{r+1})$ where $u^{\mathrm{s}} = r$ and $u^{\mathrm{w}}_{r+1} = b-1$, then

$$(Y^{\mathrm{s}}_{m+1}, Y^{\mathrm{w}}_{1,m+1}, \dots, Y^{\mathrm{w}}_{r,m+1}, Y^{\mathrm{w}}_{r+1,m+1}) = \begin{cases} \boldsymbol{\alpha}_r & \text{if } Z_{d,m} = u^{\mathrm{s}} = r, \\ (u^{\mathrm{s}}, 0, \dots, 0, u^{\mathrm{w}}_{r+1}) & \text{otherwise.} \end{cases}$$

Assigning Transition Probabilities for the Waiting-Time Transition Probability Matrix

For $m = 0, 1, 2, \cdots$, explicitly the one step transition probabilities, $p_{\boldsymbol{uv}}$ for determining the waiting-time distribution can be summarized as in the following.

If $Y_r(t_0 + m\Delta t) = (u^{\mathrm{s}}, 0, \dots, 0, u^{\mathrm{w}}_{r+1})$ where $u^{\mathrm{s}} = r$ and $B_m < b-1$, then

$$p_{\boldsymbol{u}\to\boldsymbol{v}\neq\boldsymbol{\alpha}_r} = \begin{cases} (1 - \mu_r\Delta t + o(\Delta t))\left(1 - \sum\limits_{i=1}^{R} \lambda_i \Delta t + o(\Delta t)\right) & \text{if } Z_{d,m} = 0 \text{ and} \\ & Z_{a,m} = 0, \\ (1 - \mu_r\Delta t + o(\Delta t))\left(\lambda_{Z_{a,m}}\Delta t + o(\Delta t)\right) & \text{if } Z_{d,m} = 0 \text{ and} \\ & Z_{a,m} > 0, \text{ for} \\ & Z_{a,m} = 1, \dots, R \end{cases}$$

or

$$p_{\boldsymbol{u}\to\boldsymbol{\alpha}_r} = \mu_r \Delta t + o(\Delta t) \quad \text{if } Z_{d,m} = u^{\mathrm{s}}.$$

If $(Y^{\mathrm{s}}_{m}, Y^{\mathrm{w}}_{1,m}, \dots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m}) = (u^{\mathrm{s}}, 0, \dots, 0, u^{\mathrm{w}}_{r+1})$ where $u^{\mathrm{s}} = r$ and $B_m = b-1$, then

$$p_{\boldsymbol{u}\to\boldsymbol{\alpha}_r} = \mu_r \Delta t + o(\Delta t) \quad \text{if } Z_{d,m} = u^{\mathrm{s}}$$

or

$$p_{\boldsymbol{u}\to\boldsymbol{u}} = \quad 1-\mu_r\Delta t + o(\Delta t) \quad \text{if } Z_{d,m}=0.$$

If $0<u^{\mathrm{s}}\leq r$, $u_r^{\mathrm{w}}>0$ and $B_m<b-1$, then

$$p_{\boldsymbol{u}\to\boldsymbol{v}\neq\boldsymbol{\alpha}_r} = \begin{cases} \left(1-\sum\limits_{i=1}^{R}\lambda_i\Delta t+o(\Delta t)\right)(1-\mu_{u^{\mathrm{s}}}\Delta t+o(\Delta t)) & \text{if (1c)} \\ \quad +(\lambda_{u^{\mathrm{s}}}\Delta t+o(\Delta t))\,(\mu_{u^{\mathrm{s}}}\Delta t+o(\Delta t)) & \\ \left(1-\sum\limits_{i=1}^{R}\lambda_i\Delta t+o(\Delta t)\right)(\mu_{u^{\mathrm{s}}}\Delta t+o(\Delta t)) & \text{if (2c)} \\ \big(\lambda_{Z_{a,m}}\Delta t+o(\Delta t)\big)\,(1-\mu_{u^{\mathrm{s}}}\Delta t+o(\Delta t)) & \text{if (3c)} \\ \big(\lambda_{Z_{a,m}}\Delta t+o(\Delta t)\big)\,(\mu_{u^{\mathrm{s}}}\Delta t+o(\Delta t)) & \text{if (4c)} \end{cases}$$

where

(1c) $Z_{a,m}=0$ and $Z_{d,m}=0$, or if $Z_{d,m}=u^{\mathrm{s}}<r$ and $Z_{a,m}=u^{\mathrm{s}}$,

(2c) $Z_{d,m}=u^{\mathrm{s}}$ and $Z_{a,m}=0$,

(3c) $Z_{d,m}=0$ and $Z_{a,m}>0$, for $Z_{a,m}=1,\ldots,R$,

(4c) $Z_{d,m}=u^{\mathrm{s}}$ and $Z_{a,m}>0$, for $Z_{a,m}=1,\ldots,R$ and $Z_{a,m}\neq u^{\mathrm{s}}$.

Else, if $u_r^{\mathrm{w}}>0$ and $B_m=b-1$, then

$$p_{\boldsymbol{u}\to\boldsymbol{v}} = \begin{cases} \mu_{u^{\mathrm{s}}}\Delta t+o(\Delta t) & \text{if } Z_{d,m}=u^{\mathrm{s}}, \\ 1-\mu_{u^{\mathrm{s}}}\Delta t+o(\Delta t) & \text{if } Z_{d,m}=0. \end{cases}$$

Since our interest is only in the LOS of the priority r patient receiving complete treatment without regarding all other patients lining up after, we may denote one single absorbing state $\boldsymbol{\alpha}_r$ for the Markov chain, where $\dim(\boldsymbol{\alpha}_r)=r+2$.

Following the procedures as in Fu and Lou [13] Chapters 2 and 3, with appropriate arrangement of Ω_r, the one-step time-homogeneous transition probability matrix can be partitioned in the form

$$M_r \;=\; \left(\begin{array}{c|c} N_r & \boldsymbol{c}_r' \\ \hline O_{1\times(\text{card}(\Omega_r)-1)} & 1 \end{array}\right)$$

where 1 corresponds to the transition probability of the transition from the absorbing state to itself. For $m = 1, 2, \cdots$, $N_r = (p_{\boldsymbol{uv}})$ is a matrix, called essential transition probability matrix, of one-step transition probabilities associated with non-absorbing states. $\boldsymbol{c}_r$ is a row vector whose entries are transition probabilities, $p_{\boldsymbol{u}\to\boldsymbol{\alpha}_r} \geq 0$, associated with non-absorbing states making a transition to the absorbing state.

Let $W(\boldsymbol{\alpha}_r)$ denote the waiting time of the first occurrence of event $\boldsymbol{\alpha}_r$. Discretising time and for any $m \geq 1$, $W(\boldsymbol{\alpha}_r) = n$ implies

$$\big\{Y_{rn} = \boldsymbol{\alpha}_r, Y_{r(n-1)} \neq \boldsymbol{\alpha}_r, \ldots, Y_{r2} \neq \boldsymbol{\alpha}_r, Y_{r1} \neq \boldsymbol{\alpha}_r, Y_{r0} \neq \boldsymbol{\alpha}_r\big\}$$

conditioning on $(Y_0^{\text{s}}, Y_{1,0}^{\text{w}}, \ldots, Y_{r,0}^{\text{w}}, Y_{r+1,0}^{\text{w}}) \neq \boldsymbol{\alpha}_r$.

Our main interest is to obtain the tail probability distribution

$$P^{(n)}_{\boldsymbol{y}\to\boldsymbol{\alpha}_r} = P(W(\boldsymbol{\alpha}_r) \leq n \mid (Y_0^{\text{s}}, Y_{1,0}^{\text{w}}, \ldots, Y_{r,0}^{\text{w}}, Y_{r+1,0}^{\text{w}}) = \boldsymbol{y}) \qquad \boldsymbol{y} \neq \boldsymbol{\alpha}_r \tag{3.2.1}$$

of a priority r patient being admitted to the emergency department at time t_0 to the time the patient receives complete treatment and departs **in less than or equal to n intervals of Δt.**

If we were not to assume a particular starting state $(Y^{\mathrm{s}}_0, Y^{\mathrm{w}}_{1,0}, \ldots, Y^{\mathrm{w}}_{r,0}, Y^{\mathrm{w}}_{r+1,0}) = (y^{\mathrm{s}}_0, y^{\mathrm{w}}_{1,0}, \ldots, y^{\mathrm{w}}_{r-1,0}, y^{\mathrm{w}}_{r,0}, y^{\mathrm{w}}_{r+1,0}) = \boldsymbol{y}$, we consider all possible initial states $(Y^{\mathrm{s}}_0, Y^{\mathrm{w}}_{1,0}, \ldots, Y^{\mathrm{w}}_{r,0}, Y^{\mathrm{w}}_{r+1,0}) \in \Omega_r \setminus \boldsymbol{\alpha}_r$, when a priority r patient is admitted to the system at time t_0, by assigning to each possible initial state an appropriate initial-state probability. We may find the waiting-time distribution of priority class r patients, for any t_0, to be of the form

$$P(W(\boldsymbol{\alpha}_r) \leq n \mid \boldsymbol{\xi}_r) = \boldsymbol{\xi}_r M_r^n \boldsymbol{c}', \tag{3.2.2}$$

where $\boldsymbol{\xi}_r$ is a row vector being the initial distribution, and $\boldsymbol{c} = (O_{1\times(\mathrm{card}(\Omega_r \setminus \boldsymbol{\alpha}_r))} : 1)$. Obviously (3.2.1) is a special case of (3.2.2) having a single probability value 1 in $\boldsymbol{\xi}_r$. Note that both (3.2.1) and (3.2.2) still should be considered as conditional, in the sense that the arrived customer is able to join the queue at time t_0, waiting time distributions.

Example 3.2.1. Recall in Example 3.1.1, suppose now our focus is to obtain the waiting-time distribution of a priority II patient who is admitted by the emergency department during some interval of time $[t_0 - \Delta t, t_0)$. The initial state distribution for the process $Y_{[2]m}(t)$ having state space $\{\boldsymbol{\alpha}_2$, $(1,0,1,0)$, $(1,1,1,0)$, $(1,0,2,0)$, $(1,0,1,1)$, $(2,0,0,0)$, $(2,0,1,0)$, $(2,0,2,0)$, $(2,0,1,1)$, $(2,0,0,2)\}$ has entries correspond to the following ergodic states

$$\{(0,0,0), (1,0,0), (1,1,0), (1,0,1), (2,0,0), (2,0,1)\} \tag{3.2.3}$$

just before the arrival of a priority II patient. The state $\boldsymbol{\alpha}_2$ denotes the absorbing state, in this example, indicating the completion of treatment of a priority II patient being admitted to the emergency department during some interval of time $[t_0 - \Delta t, t_0)$.

We keep $\hat{\pi}_k$, the kth entry of $\hat{\boldsymbol{\pi}}$, for all k's corresponding to the states in (3.2.3), and assign $\hat{\pi}_{k'} = 0$ for all k''s corresponding to the states not in (3.2.3). We then normalize the non-zero entries such that the sum then is 1. Therefore, the initial distribution, denoted by $\boldsymbol{\xi}_2$, for obtaining the waiting-time distribution of the entry of a priority II patient is

$$\boldsymbol{\xi}_2 = (0, 0.2051, 0.0440, 0.0097, 0, 0.1090, 0.2912, 0.3409, 0, 0)$$

with non-zero entries correspond to the initial states $(1, 0, 1, 0)$, $(1, 1, 1, 0)$, $(1, 0, 2, 0)$, $(2, 0, 0, 0)$, $(2, 0, 1, 0)$ and $(2, 0, 2, 0)$.

□

Suppose one is interested in the waiting-time probability distribution of a priority r patient from the time of admission to an emergency department at time t_0 to the time the patient departs **in exactly n intervals of time**. Applying Theorem 2.2 in Fu and Lou [13], we have

$$P\left(W(\boldsymbol{\alpha}_r) = n \mid \boldsymbol{\xi}_r\right) = \boldsymbol{\xi}_r N_r^{n-1}(I - N_r)\mathbf{1}' \tag{3.2.4}$$

where I is an identity matrix having the same dimension of N_r and $\mathbf{1}$ is a unit row vector of proper length.

3.2.3 Deriving the Expectation of Wait Time

By the definition of a generating function as in Feller [9] Chapter XI, without duplication and confusion but in the context of our problem, we will use notations similar to those in Fu and Lou [13] Chapters 3 and 5. The probability generating function of $W(\boldsymbol{\alpha}_r)$, denoted by $\varphi_W(s)$ and by definition, can be written as

$$\varphi_W(s) = \sum_{k=0}^{\infty} s^k \Pr\left\{W(\boldsymbol{\alpha}_r) = k \mid \boldsymbol{\xi}_r\right\} = \sum_{k=1}^{\infty} s^k \Pr\left\{W(\boldsymbol{\alpha}_r) = k \mid \boldsymbol{\xi}_r\right\} \qquad (3.2.5)$$

if we assume $\Pr\{W(\boldsymbol{\alpha}_r) = 0\} = 0$. By this definition, it is not difficult to see that if we differentiate $\varphi_W(s)$ once and evaluate the function at $s = 1$, we will obtain

$$\varphi_W^{(1)}(s)\big|_{s=1} = \sum_{k=1}^{\infty} k \Pr\{W(\boldsymbol{\alpha}_r) = k \mid \boldsymbol{\xi}_r\} = E[W(\boldsymbol{\alpha}_r)] \qquad (3.2.6)$$

which is the expected waiting-time of a priority i patient in an emergency department since he or she being admitted at some time t_0 to the time of his or her departure. Note that in general, the assumption $\Pr\{W(\boldsymbol{\alpha}_r) = 0\} = 0$ is not needed for the derivation of (3.2.6) when the definition of $W(\boldsymbol{\alpha}_r)$ is other than the LOS we defined to be in this section.

Fu and Lou [13] in Section 5.4 clearly proved that (3.2.5) and (3.2.6) can be expressed in terms of $\boldsymbol{\xi}_r$ and N_r as

$$\varphi_W(s) = 1 + (s-1)\boldsymbol{\xi}_r(I - sN_r)^{-1}\mathbf{1}', \qquad (3.2.7)$$

and

$$E[W(\boldsymbol{\alpha}_r)] = \boldsymbol{\xi}_r(I - N_r)^{-1}\mathbf{1}'. \qquad (3.2.8)$$

3.2.4 Deriving the Variance of Waiting-Time

Commonly, the variance of a random variable Z is calculated using the formula $\mathrm{Var}(Z) = E[Z^2] - (E[Z])^2$. Since we have an expression for $E[W(\boldsymbol{\alpha}_r)]$, an expression for $E[W^2(\boldsymbol{\alpha}_r)]$ be determined enables us to compute the variance of $W(\boldsymbol{\alpha}_r)$. We will show that

$$E[W^2(\boldsymbol{\alpha}_r)] = \boldsymbol{\xi}_r(I - N_r)^{-1}(I + N_r)(I - N_r)^{-1}\mathbf{1}' \tag{3.2.9}$$

Alternatively, Fu, Spiring and Xie in [14] Theorem 1 (iii) gives an expression of $E[W^2(\boldsymbol{\alpha}_r)]$ as $\boldsymbol{\xi}_r(I + N_r)(I - N_r)^{-2}\mathbf{1}'$ and we will show that this expression and (3.2.9) are equivalent. First, we attempt to show the derivation of $E[W^2(\boldsymbol{\alpha}_r)]$ in [14].

Differentiating $\varphi_W(s)$, by its definition as in (3.2.5), twice and evaluating the function at $s = 1$ yields

$$\begin{aligned} \varphi_W^{(2)}(s)\big|_{s=1} &= \sum_{k=2}^{\infty} k(k-1)\Pr\{W(\boldsymbol{\alpha}_r) = k \mid \boldsymbol{\xi}_r\} \\ &= \sum_{n=1}^{\infty} k(k-1)\Pr\{W(\boldsymbol{\alpha}_r) = k \mid \boldsymbol{\xi}_r\} \\ &= E\left[W^2(\boldsymbol{\alpha}_r)\right] - E\left[W(\boldsymbol{\alpha}_r)\right] \end{aligned} \tag{3.2.10}$$

By (3.2.5), $\varphi_W^{(2)}(s)$ can also be expressed as

$$\begin{aligned} \varphi_W^{(2)}(s) &= \sum_{k=2}^{\infty} k(k-1)s^{k-2}\boldsymbol{\xi}_r N_r^{k-1}(I - N_r)\mathbf{1}' \\ &= \boldsymbol{\xi}_r N_r \left(\sum_{k=2}^{\infty} k(k-1)s^{k-2}N_r^{k-2}\right)(I - N_r)\mathbf{1}' \\ &= 2\,\boldsymbol{\xi}_r N_r(I - sN_r)^{-3}(I - N_r)\mathbf{1}' \end{aligned}$$

Therefore,

$$\begin{aligned} E[W^2(\boldsymbol{\alpha}_r)] &= \varphi_W^{(2)}(s)\big|_{s=1} + E[W(\boldsymbol{\alpha}_r)] \\ &= 2\,\boldsymbol{\xi}_r N_r(I-N_r)^{-2}\mathbf{1}' + \boldsymbol{\xi}_r(I-N_r)^{-1}\mathbf{1}' \\ &= \boldsymbol{\xi}_r(I+N_r)(I-N_r)^{-2}\mathbf{1}' \qquad (3.2.11) \end{aligned}$$

Our method to derive the form of $\varphi_W^{(2)}(s)$ in product of matrices is by differentiating equation (3.2.7) twice and evaluating the expression at the value $s = 1$,

$$\begin{aligned} \frac{d\varphi_W(s)}{ds} &= \boldsymbol{\xi}_r(I-sN_r)^{-1}\mathbf{1}' + (s-1)\frac{d}{ds}\left[\boldsymbol{\xi}_r(I-sN_r)^{-1}\mathbf{1}'\right] \\ \frac{d^2\varphi_W(s)}{(ds)^2} &= \frac{d}{ds}\left[\boldsymbol{\xi}_r(I-sN_r)^{-1}\mathbf{1}'\right] + \frac{d}{ds}\left[\boldsymbol{\xi}_r(I-sN_r)^{-1}\mathbf{1}'\right] \\ &\quad + (s-1)\frac{d}{ds}\left(\frac{d}{ds}\left[\boldsymbol{\xi}_r(I-sN_r)^{-1}\mathbf{1}'\right]\right) \qquad (3.2.12) \end{aligned}$$

Evaluating (3.2.12) at $s = 1$ is to evaluate

$$2\,\frac{d}{ds}\left[\boldsymbol{\xi}_r(I-sN_r)^{-1}\mathbf{1}'\right]$$

at $s = 1$. The main term to look for in the above is $\frac{d}{ds}[\boldsymbol{\xi}_r(I-sN_r)^{-1}\mathbf{1}']$. Let $A^{-1}(s) = (I-sN_r)^{-1}$ and using the basic rule of differentiating the inverse of a matrix

$$\frac{d}{ds}A^{-1}(s) = -A^{-1}(s)\left[\frac{d}{ds}A(s)\right]A^{-1}(s)$$

(see [26]). It is not difficult to show

$$\left.\frac{d^2\varphi_W(s)}{(ds)^2}\right|_{s=1} = 2\,\boldsymbol{\xi}_r(I-N_r)^{-1}N_r(I-N_r)^{-1}\mathbf{1}' \qquad (3.2.13)$$

From (3.2.10) and (3.2.13), we have

$$\begin{aligned} E[W^2(\boldsymbol{\alpha}_r)] =&\ \varphi_W^{(2)}(s)\big|_{s=1} + E[W(\boldsymbol{\alpha}_r)] \\ =&\ 2\,\boldsymbol{\xi}_r(I-N_r)^{-1}N_r(I-N_r)^{-1}\mathbf{1}' + \boldsymbol{\xi}_r(I-N_r)^{-1}\mathbf{1}' \qquad (3.2.14) \\ =&\ \boldsymbol{\xi}_r(I-N_r)^{-1}(I+N_r)(I-N_r)^{-1}\mathbf{1}' \end{aligned}$$

To show that (3.2.11) and the last line of (3.2.14) are equivalent, it suffices to show that $(I-N_r)^{-1}$ and $(I+N_r)$ commute. First, since

$$\begin{aligned} (I-N_r)(I+N_r) &= I - N_r + N_r - N_rN_r \\ &= N_r - N_rN_r + I - A \\ &= N_r(I-N_r) + (I-N_r) \\ &= (I+N_r)(I-N_r), \end{aligned}$$

we have $(I+N_r)$ and $(I-N_r)$ commute. By Proposition 3.3.1, $(I-N_r)^{-1}$ and $(I+N_r)$ commute.

Proposition 3.3.1. *If two square matrices A and B commutate and A is invertible, then A^{-1} and B also commute.*

Proof. Since A and B commute and A is invertible, we can write

$$\begin{aligned} AB &= BA \\ A^{-1}ABA^{-1} &= A^{-1}BAA^{-1} \\ BA^{-1} &= A^{-1}B. \end{aligned}$$

Thanks to Dr. Brad Johnson at the Department of Statistics, University of Manitoba, for the idea and discussion of this simple proof. □

With the form of $E\left[W^2(\boldsymbol{\alpha}_r)\right] = \boldsymbol{\xi}_r(I-N_r)^{-1}(I+N_r)(I-N_r)^{-1}\mathbf{1}'$, the variance of the random variable $W(\boldsymbol{\alpha}_r)$ is

$$\begin{aligned}\mathrm{Var}\left(W(\boldsymbol{\alpha}_r)\right) &= E\left[W^2(\boldsymbol{\alpha}_r)\right] - \left(E[W(\boldsymbol{\alpha}_r)]\right)^2 \\ &= \boldsymbol{\xi}_r(I-N_r)^{-1}(I+N_r)(I-N_r)^{-1}\mathbf{1}' - [\boldsymbol{\xi}_r(I-N_r)^{-1}\mathbf{1}']^2 \\ &= \boldsymbol{\xi}_r(I-N_r)^{-1}(I+N_r-\mathbf{1}'\boldsymbol{\xi}_r)(I-N_r)^{-1}\mathbf{1}'\end{aligned}$$

3.2.5 Re-defining the Variable *Wait Time*

If our interest is to obtain the distribution of the re-defined variable **wait time**: starting from the point a customer joined the queueing system to the time service begins on the same customer for the *first time*. This is where the FMCI technique glows as only a minor modification on the imbedding and the state space will allow us to obtain the distribution of the re-defined $W(\boldsymbol{\alpha}_r)$. We will illustrate the procedure on preemptive priority queues directly in determining the transition rules and probabilities of the process.

Finding the Exact Tail Distribution of *Wait time*

Conditioning on that a customer of service priority r $(1 \leq r \leq R)$ arrived and is admitted to the queueing system during $[t_0 - \Delta t, t_0)$ for some t_0. Re-defining the state transition rules and transition probabilities is necessary in order to calculate the tail probabilities of wait time. Let $Y_{rm} = (Y^{\mathrm{s}}_m, Y^{\mathrm{w}}_{1,m}, \ldots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m})$ be as previously defined and let $W(\boldsymbol{\alpha}_r)$ denote the waiting time of the event $\boldsymbol{\alpha}_r$ where $\boldsymbol{\alpha}_r \in \Omega_r$ is a vector used to denote the absorbing state indicating that service begins for the first time on the priority r customer of interest. If during $[t_0 - \Delta t, t_0)$ the queueing system ever becomes empty or that a service priority $j > r$ customer is in service, then $Y_{r0} = \boldsymbol{\alpha}_r$ and $W(\boldsymbol{\alpha}_r) = 0$ trivially under the preemptive discipline. Otherwise, discretising time and for any $n \geq 1$, $W(\boldsymbol{\alpha}_r) = n$ implies

$$\left\{ Y_{rn} = \boldsymbol{\alpha}_r, Y_{r(n-1)} \neq \boldsymbol{\alpha}_r, \ldots, Y_{r1} \neq \boldsymbol{\alpha}_r, Y_{r0} \neq \boldsymbol{\alpha}_r \right\}.$$

With the new definition of $\boldsymbol{\alpha}_r$ and $W(\boldsymbol{\alpha}_r)$, the newly induced state space of the process Y_{rm} consists of collection of states which satisfies the following criteria:

$$\begin{aligned} \Omega_r = \ & \{\boldsymbol{\alpha}_r\} \bigcup \{(y^{\mathrm{s}}, y_1^{\mathrm{w}}, \ldots, y_r^{\mathrm{w}}) \mid 1 \leq y^{\mathrm{s}} \leq r, \ 0 \leq y_1^{\mathrm{w}} \leq b-1, \\ & \quad 0 \leq y_2^{\mathrm{w}} \leq b - 1 - y_1^{\mathrm{w}}, \ 0 \leq y_i^{\mathrm{w}} \leq b - 1 - \sum_{j=1}^{i-1} y_j^{\mathrm{w}} \\ & \quad \text{for } i = 3, \ldots, (r-1), \ 1 \leq y_r^{\mathrm{w}} \leq b - 1 - \sum_{j=1}^{r-1} y_j^{\mathrm{w}}, \\ & \quad 0 \leq y_{r+1}^{\mathrm{w}} \leq b - 1 - \sum_{j=1}^{r} y_j^{\mathrm{w}} \text{ and for each } (y^{\mathrm{s}}, y_1^{\mathrm{w}}, \ldots, y_r^{\mathrm{w}}), \\ & \quad y_j^{\mathrm{w}} = 0 \text{ for all } j = 1, \ldots, y^{\mathrm{s}} - 1 \text{ when } y^{\mathrm{s}} > 1\}. \end{aligned}$$

For the convenience of stating the state transition rules and state transition probabilities, suppose we are given $(Z_{a,m}, Z_{d,m})$ and $(Y_m^{\mathrm{s}}, Y_{1,m}^{\mathrm{w}}, \ldots, Y_{r,m}^{\mathrm{w}}, Y_{r+1,m}^{\mathrm{w}}) = (u^{\mathrm{s}}, u_1^{\mathrm{w}}, \ldots, u_r^{\mathrm{w}}, \ u_{r+1}^{\mathrm{w}}) = \boldsymbol{u}$ and the process will make a transition to a state $(Y_{m+1}^{\mathrm{s}}, Y_{1,m+1}^{\mathrm{w}}, \ldots, Y_{r,m+1}^{\mathrm{w}}, \ Y_{r+1,m+1}^{\mathrm{w}}) = (v^{\mathrm{s}}, v_1^{\mathrm{w}}, \ldots, v_r^{\mathrm{w}}, v_{r+1}^{\mathrm{w}}) = \boldsymbol{v}$ for any $m \in \mathbb{N}_0$.

Observe that by definition, from t_0 and for subsequent moments $t_0 + m\Delta t, m = 1, 2, \cdots$, $Y_{r+1,m}^{\mathrm{w}}$ is non-decreasing and $Y_m^{\mathrm{s}} = r$ only if $Y_{i,m}^{\mathrm{w}} = 0$ for all $i = 1, \ldots, r-1$. When the system reaches the state $(Y_m^{\mathrm{s}} = r, 0, \ldots, 0, Y_{r+1,m}^{\mathrm{w}})$ at some m for the *first* time, the system has entered the absorbing state $\boldsymbol{\alpha}_r$. Our main interest can be properly formulated as to obtaining the tail probability distribution of the variable $W(\boldsymbol{\alpha}_r)$

$$P^{(n)}_{(Y_0^{\rm s},Y_{1,0}^{\rm w},\ldots,Y_{r,0}^{\rm w},Y_{r+1,0}^{\rm w})\to\boldsymbol{\alpha}_r} = P(W(\boldsymbol{\alpha}_r) \le n\Delta t \mid \boldsymbol{\xi}_r) \tag{3.2.15}$$

such that the absorbing state is reached for the first time **at or after** time t_0. It is the probability that given a service priority r customer be admitted to the system at time t_0, this customer **waits less than or equal to n increments of** Δt, $n = 0, 1, 2, \cdots$, until service begins for the same priority r customer for first time since the customer joined the system.

If a particular priority r customer of interest arrived during $[t_0 - \Delta t, t_0)$ and begins to receive service at time t_0, then the state $\boldsymbol{\alpha}_r$ is reached. However, we want to closely examine the case when the particular priority r customer of interest waits in line for service after arriving to the system. We give the state transition rules and transition probabilities as in the following.

State Transition Rules

For $m = 0, 1, 2, \cdots$, if $0 < u^{\rm s} \le r$, $u_i^{\rm w} = 0$ for $i = 1, \ldots, r-1$, $u_r^{\rm w} = 1$, and $B_m < b$, then

$$v^{\rm s} = \begin{cases} \min(Z_{a,m}, u^{\rm s}) & \text{if } Z_{d,m} = 0 \text{ and } 0 < Z_{a,m} < r, \\ Z_{a,m} & \text{if } Z_{d,m} = u^{\rm s} \text{ and } 0 < Z_{a,m} < r, \\ \alpha^{\rm s} & \text{if } Z_{d,m} = u^{\rm s} \text{ and, } Z_{a,m} = 0 \text{ or } Z_{a,m} \ge r, \\ u^{\rm s} & \text{otherwise,} \end{cases}$$

for $\ell = 1, \ldots, r-1$,

$$v_\ell^{\mathrm{w}} = \begin{cases} 1 & \text{if } Z_{d,m} = 0 \text{ and } 0 < Z_{a,m} < r \text{ and } \ell = \max(Z_{a,m}, u^{\mathrm{s}}), \\ \alpha_\ell^{\mathrm{w}} & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and, } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ u_\ell^{\mathrm{w}} & \text{otherwise,} \end{cases}$$

$$v_r^{\mathrm{w}} = \begin{cases} 2 & \text{if } u^{\mathrm{s}} = r,\ Z_{d,m} = 0 \text{ and } 0 < Z_{a,m} < r, \\ \alpha_r^{\mathrm{w}} & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and, } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ 1 & \text{otherwise,} \end{cases}$$

$$v_{r+1}^{\mathrm{w}} = \begin{cases} u_{r+1}^{\mathrm{w}} + 1 & \text{if } Z_{d,m} = 0 \text{ and } Z_{a,m} \geq r, \\ \alpha_{r+1}^{\mathrm{w}} & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and, } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ u_{r+1}^{\mathrm{w}} & \text{otherwise,} \end{cases}$$

where α^{s} and α_ℓ^{w} for $\ell = 1, \ldots, r+1$ are the components of the absorbing state $\boldsymbol{\alpha}_r = (\alpha^{\mathrm{s}}, \alpha_1^{\mathrm{w}}, \ldots, \alpha_r^{\mathrm{w}}, \alpha_{r+1}^{\mathrm{w}})$.

For $m = 1, 2, \cdots$, if $0 < u^{\mathrm{s}} \leq r$, $u_r^{\mathrm{w}} > 1$ or $u_r^{\mathrm{w}} > 0$ and $u_i^{\mathrm{w}} > 0$ for some $i \in \{1, \ldots, r-1\}$ and $B_m < b$, then

$$v^{\mathrm{s}} = \begin{cases} \min(Z_{a,m}, u^{\mathrm{s}}) & \text{if } Z_{d,m} = 0 \text{ and } 0 < Z_{a,m} < r, \\ \min(Z_{a,m}, K_m^*) & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and } 0 < Z_{a,m} < r, \\ K_m^* & \text{if } Z_{d,m} = u^{\mathrm{s}} \text{ and, } Z_{a,m} = 0 \text{ or } Z_{a,m} > K_m^*, \\ u^{\mathrm{s}} & \text{otherwise,} \end{cases}$$

for $\ell = 1, \ldots, r-1$,

$$v^{\mathrm{w}}_{\ell} = \begin{cases} u^{\mathrm{w}}_{\ell} + 1 & \text{if } Z_{d,m} = 0,\ 0 < Z_{a,m} < r \text{ and } \ell = \max(Z_{a,m}, u^{\mathrm{s}}), \\ u^{\mathrm{w}}_{\ell} - 1 & \text{if } Z_{d,m} = u^{\mathrm{s}},\ Z_{a,m} = 0 \text{ or } z_a > K^*_m, \text{ and } \ell = K^*_m, \\ u^{\mathrm{w}}_{\ell} & \text{otherwise,} \end{cases}$$

$$v^{\mathrm{w}}_{r} = \begin{cases} u^{\mathrm{w}}_{r} + 1 & \text{if } u^{\mathrm{s}} = r,\ Z_{d,m} = 0 \text{ and } 0 < Z_{a,m} < r, \\ u^{\mathrm{w}}_{r} - 1 & \text{if } u^{\mathrm{w}}_{r} > 1,\ Z_{d,m} = u^{\mathrm{s}} \text{ and, } Z_{a,m} = 0 \text{ or } Z_{a,m} \geq r, \\ u^{\mathrm{w}}_{r} & \text{otherwise,} \end{cases}$$

$$v^{\mathrm{w}}_{r+1} = \begin{cases} u^{\mathrm{w}}_{r+1} + 1 & \text{if } Z_{a,m} \geq r, \\ u^{\mathrm{w}}_{r+1} & \text{otherwise.} \end{cases}$$

For $m = 1, 2, \cdots$, if $B_m = b$, then $Z_{a,m} = 0$ and

$$v^{\mathrm{s}} = \begin{cases} K^*_m & \text{if } Z_{d,m} = u^{\mathrm{s}},\ u^{\mathrm{w}}_{r} > 1 \text{ or, } u^{\mathrm{w}}_{r} > 0 \text{ and } u^{\mathrm{w}}_{k} > 0 \text{ for} \\ & \text{some } k \in \{1, \ldots, r-1\}, \\ \alpha^{\mathrm{s}} & \text{if } Z_{d,m} = u^{\mathrm{s}},\ u^{\mathrm{w}}_{r} = 1 \text{ and } u^{\mathrm{w}}_{k} = 0 \text{ for } k = 1, \ldots, r-1, \\ u^{\mathrm{s}} & \text{otherwise,} \end{cases}$$

for $\ell = 1, \ldots, r-1$,

$$v^{\mathrm{w}}_{\ell} = \begin{cases} u^{\mathrm{w}}_{\ell} - 1 & \text{if } Z_{d,m} = u^{\mathrm{s}},\ u^{\mathrm{w}}_{r} > 1, \text{ or } u^{\mathrm{w}}_{r} > 0 \text{ and } u^{\mathrm{w}}_{k} > 0 \\ & \text{for some } k \in \{1, \ldots, r-1\}, \text{ and } \ell = K^*_m, \\ \alpha^{\mathrm{w}}_{\ell} & \text{if } Z_{d,m} = u^{\mathrm{s}},\ u^{\mathrm{w}}_{r} = 1 \text{ and } u^{\mathrm{w}}_{k} = 0 \text{ for all} \\ & k = 1, \ldots, r-1, \\ u^{\mathrm{w}}_{\ell} & \text{otherwise,} \end{cases}$$

$$v_r^{\mathrm{w}} = \begin{cases} u_r^{\mathrm{w}} - 1 & \text{if } u_r^{\mathrm{w}} > 1,\ Z_{d,m} = u^{\mathrm{s}} \text{ and } u_k^{\mathrm{w}} = 0 \text{ for } k = 1, \ldots, r-1, \\ \alpha_r^{\mathrm{w}} & \text{if } Z_{d,m} = u^{\mathrm{s}},\ u_r^{\mathrm{w}} = 1 \text{ and } u_k^{\mathrm{w}} = 0 \text{ for } k = 1, \ldots, r-1, \\ u_r^{\mathrm{w}} & \text{otherwise,} \end{cases}$$

$$v_{r+1}^{\mathrm{w}} = \begin{cases} \alpha_{r+1}^{\mathrm{w}} & \text{if } Z_{d,m} = u^{\mathrm{s}},\ u_r^{\mathrm{w}} = 1 \text{ and } u_k^{\mathrm{w}} = 0 \text{ for } k = 1, \ldots, r-1, \\ u_{r+1}^{\mathrm{w}} & \text{otherwise.} \end{cases}$$

State Transition Probabilities

For $m = 0, 1, 2, \cdots$, the one-step transition probabilities for determining the waiting-time distribution can be summarized as in the following.

Assuming that the service priority r customer is waiting in queue and the queue capacity is not reached, if $u_r^{\mathrm{w}} = 1$, $u_i^{\mathrm{w}} = 0$ for $i = 1, \ldots, r-1$, then $Z_{d,m} = u^{\mathrm{s}}$ and, $Z_{a,m} = 0$ or $Z_{a,m} \geq r$ with probability

$$p_{\boldsymbol{u} \to \boldsymbol{\alpha}_r} = \left(1 - \sum_{i=1}^{r-1} \lambda_i \Delta t + o(\Delta t)\right) \left(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)\right)$$

or

$$p_{\boldsymbol{u} \to \boldsymbol{v} \neq \boldsymbol{\alpha}_r} = \begin{cases} \left(1 - \sum_{i=1}^{r} \lambda_i \Delta t + o(\Delta t)\right) \left(1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)\right) & \text{if (1d)} \\ \qquad + \left(\lambda_{u^{\mathrm{s}}} \Delta t + o(\Delta t)\right) \left(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)\right) & \\ \left(\lambda_{Z_{a,m}} \Delta t + o(\Delta t)\right) \left(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)\right) & \text{if (2d)} \\ \left(\lambda_{Z_{a,m}} \Delta t + o(\Delta t)\right) \left(1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)\right) & \text{if (3d)} \end{cases}$$

where

(1d) $Z_{d,m} = 0$ and $Z_{a,m} = 0$, or $Z_{d,m} = u^{\mathrm{s}} < r$ and $Z_{a,m} = u^{\mathrm{s}}$, (2d) $Z_{d,m} = u^{\mathrm{s}}$ and $Z_{a,m} > 0$ for $Z_{a,m} = 1, \ldots, r-1$, $Z_{a,m} \neq u^{\mathrm{s}}$,

(3d) $Z_{d,m} = 0$ and $Z_{a,m} > 0$ for $Z_{a,m} = 1, \ldots, r$.

If $u_r^{\mathrm{w}} > 1$ or $u_i^{\mathrm{w}} > 0$ for some $i \in \{1, \ldots, r-1\}$, then

$$p_{\boldsymbol{u} \to \boldsymbol{v} \neq \boldsymbol{\alpha}_r} = \begin{cases} \left(1 - \sum\limits_{i=1}^{r} \lambda_i \Delta t + o(\Delta t)\right)(1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (1e)} \\ \qquad + (\lambda_{u^{\mathrm{s}}} \Delta t + o(\Delta t))(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \\ \left(1 - \sum\limits_{i=1}^{r} \lambda_i \Delta t + o(\Delta t)\right)(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (2e)} \\ \left(\lambda_{Z_{a,m}} \Delta t + o(\Delta t)\right)(1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (3e)} \\ \left(\lambda_{Z_{a,m}} \Delta t + o(\Delta t)\right)(\mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)) & \text{if (4e)} \end{cases}$$

where

(1e) $Z_{d,m} = 0$ and $Z_{a,m} = 0$, or if $Z_{d,m} = u^{\mathrm{s}} < r$ and $Z_{a,m} = u^{\mathrm{s}}$,

(2e) $Z_{d,m} = u^{\mathrm{s}}$ and $Z_{a,m} = 0$,

(3e) $Z_{d,m} = 0$ and $Z_{a,m} > 0$, for $Z_{a,m} = 1, \ldots, r$,

(4e) $Z_{d,m} = u^{\mathrm{s}}$ and $Z_{a,m} > 0$ for $Z_{a,m} = 1, \ldots, r$, $Z_{a,m} \neq u^{\mathrm{s}}$.

Assuming that the service priority r customer is waiting in queue and $B_m = b$, if $u_r^{\mathrm{w}} = 1$, $u_i^{\mathrm{w}} = 0$ for $i = 1, \ldots, r-1$, then $Z_{d,m} = u^{\mathrm{s}}$ with probability

$$p_{\boldsymbol{u} \to \boldsymbol{\alpha}_r} = \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t)$$

or $Z_{d,m} = 0$ with probability

$$p_{\boldsymbol{u} \to \boldsymbol{u}} = 1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t).$$

Else, if $u_r^{\mathrm{w}} > 1$ or $u_k^{\mathrm{w}} > 0$ for some $k \in \{1, \ldots, r-1\}$, then

$$p_{\boldsymbol{u} \to \boldsymbol{v} \neq \boldsymbol{\alpha}_r} = \begin{cases} \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t) & \text{if } Z_{d,m} = u^{\mathrm{s}}, \\ \\ 1 - \mu_{u^{\mathrm{s}}} \Delta t + o(\Delta t) & \text{if } Z_{d,m} = 0. \end{cases}$$

Similarly, with appropriate arrangement of Ω_r, the one-step time homogeneous tpm M_r can be partitioned as

$$M_r = \left(\begin{array}{c|c} N_r & \boldsymbol{c}_r' \\ \hline O_{1 \times (\mathrm{card}(\Omega_r)-1)} & 1 \end{array} \right). \tag{3.2.16}$$

Considering all possible initial status $(Y_0^{\mathrm{s}}, Y_{1,0}^{\mathrm{w}}, \ldots, Y_{r,0}^{\mathrm{w}}, Y_{r+1,0}^{\mathrm{w}}) \in \Omega_r$ at time t_0 and assigning to each possible starting state an appropriate initial-state probability such that $\sum\limits_{\boldsymbol{y} \in \Omega_r} \Pr\{(Y_0^{\mathrm{s}}, Y_{1,0}^{\mathrm{w}}, \ldots, Y_{r,0}^{\mathrm{w}}, Y_{r+1,0}^{\mathrm{w}}) = \boldsymbol{y}\} = 1$. The general waiting-time distribution of a priority r customer, for any t_0, now is of the form

$$\begin{aligned} & P(W(\boldsymbol{\alpha}_r) \leq n\Delta t \mid \boldsymbol{\xi}_r) \\ & = \sum_{\boldsymbol{y} \in \Omega_r} \Pr\{(Y_0^{\mathrm{s}}, Y_{1,0}^{\mathrm{w}}, \ldots, Y_{r,0}^{\mathrm{w}}, Y_{r+1,0}^{\mathrm{w}}) = \boldsymbol{y}\} \cdot P_{\boldsymbol{y} \to \boldsymbol{\alpha}_r}^{(n)} \end{aligned} \tag{3.2.17}$$

with $\lim\limits_{n \to \infty} \sum\limits_{\boldsymbol{y} \in \Omega_r} \Pr\{(Y_0^{\mathrm{s}}, Y_{1,0}^{\mathrm{w}}, \ldots, Y_{r,0}^{\mathrm{w}}, Y_{r+1,0}^{\mathrm{w}}) = \boldsymbol{y}\} \cdot P_{\boldsymbol{y} \to \boldsymbol{\alpha}_r}^{(n)} = 1$.

Equation (3.2.17) can be expressed as a product

$$\boldsymbol{\xi}_r M_r^n \boldsymbol{c}', \tag{3.2.18}$$

where $\boldsymbol{\xi}_r$ is a row vector being the initial-state distribution having non-negative entries summing to one, and $\boldsymbol{c} = (\mathbf{0}_{1 \times (\mathrm{card}(\Omega_r - 1))} : 1)$. Obviously (3.2.15) is a special case of (3.2.17) which has a single probability value 1 in $\boldsymbol{\xi}_r$. To obtain the first and second moments of $W(\boldsymbol{\alpha}_r)$, equations (3.2.8) and (3.2.14) also can be applied here.

Example 3.2.2. Continuing from Example 3.1.1, from the ergodic states an urgent patient can only be admitted into the system in the following possible states

$$\{(0,0,0),(1,0,0),(1,1,0),(1,0,1),(2,0,0)\}.$$

$Y_{[1]m}$ has a state space

$$\Omega_1 = \{(1,1,0),(1,2,0),(1,1,1),\boldsymbol{\alpha}_1\}$$

and a normalized initial distribution

$$\boldsymbol{\xi}_1 = (0.0562, 0.0125, 0.1059, 0.8255). \tag{3.2.19}$$

The expectation of wait time of an urgent patient admitted to ED is 14.96 minutes and s.d. is 48.11 minutes.

If a non-urgent type patient is admitted and has to wait, $Y_{[2]m}$ has a state space

$$\Omega_2 = \left\{ \begin{array}{c} (1,0,1,0),(2,0,1,0),(1,1,1,0),(1,0,2,0), \\ (2,0,2,0),(1,0,1,1),(2,0,1,1),\boldsymbol{\alpha}_2 \end{array} \right\}$$

and a normalized initial distribution

$$\boldsymbol{\xi}_2 = (0.0562, 0.2670, 0.0125, 0.1059, 0.2800, 0, 0, 0.2785). \tag{3.2.20}$$

The expectation of wait time of a non-urgent patient admitted to ED is 51.81 minutes and s.d. is 74.51 minutes.

Chapter 4

Single-Server Priority Queues with Service Thresholds

In this chapter, we will demonstrate using the same technique to model $M/M/1$-$R/b/c$ preemptive and non-preemptive priority queues. Now, instead of allowing only one customer at a time to be in service, we allow at the same time up to c to be in service. Then we show the procedure for obtaining the probability distribution of the variable **wait time** - starting from the time a customer joins the queueing system to the time of receiving service for the first time.

4.1 Preemptive Model

Detailed description of a preemptive priority queue is discussed in Section §1.3 and in Chapter 3. In this section, we continue to assume the preemptive repeat-different discipline where a preempted customer keeps no memory of previous service done and resumes service afresh after returning to service from preemptions. Some application of this model can be seen in a hospital emergency department where a medical team can provide treatment to, whether a physician was on site or was temporarily off site waiting for new diagnostic reports on patients, a number of patients at the same time up to some threshold.

Now, with the parameter $c > 1$, the states used to describe the status of the priority queue requires modification. We will use a vector $X_m = (X^{\mathrm{s}}_{1,m}, \ldots, X^{\mathrm{s}}_{R,m}, X^{\mathrm{w}}_{1,m}, \ldots, X^{\mathrm{w}}_{R,m})$ where at time points $m\Delta t$ for $m = 0, 1, 2, \cdots$, $X^{\mathrm{s}}_{i,m}$ monitors the number of priority i customers in service and $X^{\mathrm{w}}_{i,m}$ monitors the number of priority i customers waiting for service for $i = 1, \ldots, R$. With a service threshold c, $1 < c < b$, and a system capacity $b < \infty$, the X_m induces a finite state space Ω_X which can be constructed according to the following criteria:

$$\begin{aligned}\Omega_X = \quad & \{\mathbf{0}\} \bigcup \, \{(x^{\mathrm{s}}_1, \ldots, x^{\mathrm{s}}_R, x^{\mathrm{w}}_1, \ldots, x^{\mathrm{w}}_R) \mid 0 \le x^{\mathrm{s}}_1 \le c, 0 \le x^{\mathrm{w}}_1 \le b - c, \\ & \qquad 0 \le x^{\mathrm{s}}_i \le c - \sum_{k=1}^{i-1} x^{\mathrm{s}}_k \text{ for } i = 2, \ldots, R, \\ & \qquad 0 \le x^{\mathrm{w}}_i \le b - c - \sum_{k=1}^{i-1} x^{\mathrm{w}}_k \text{ for } i = 2, \ldots, R, \\ & \qquad \text{and for each } (x^{\mathrm{s}}_1, \ldots, x^{\mathrm{s}}_R, x^{\mathrm{w}}_1, \ldots, x^{\mathrm{w}}_R),\ x^{\mathrm{w}}_i = 0 \text{ for} \\ & \qquad i = 1, \ldots, k^* - 1\}\end{aligned}$$

where $k^* = \max\{k : x^{\mathrm{s}}_k > 0 \text{ for } k = 1, \ldots, R \mid (x^{\mathrm{s}}_1, \ldots, x^{\mathrm{s}}_R, x^{\mathrm{w}}_1, \ldots, x^{\mathrm{w}}_R) \neq \mathbf{0}\}$.

Let $B_m = \sum_{i=1}^{R}(X^{\mathrm{s}}_{i,m} + X^{\mathrm{w}}_{i,m})$ and $C_m = \sum_{i=1}^{R} X^{\mathrm{s}}_{i,m}$ be the number of customers in the system and the number of customers in service, respectively, at time $m\Delta t$ for

$m = 0, 1, \cdots$. For the convenience when describing a one-step transition probability by $p_{\boldsymbol{uv}} = \Pr\{X_{m+1} = \boldsymbol{v} \mid X_m = \boldsymbol{u}\}$ for any m, we use the notation $\boldsymbol{u} \to \boldsymbol{v}$ to describe a one-step state transition of the process going from $(X^{\mathrm{s}}_{1,m}, \ldots, X^{\mathrm{s}}_{R,m}, X^{\mathrm{w}}_{1,m}, \ldots, X^{\mathrm{w}}_{R,m}) = (u^{\mathrm{s}}_1, \ldots, u^{\mathrm{s}}_R, u^{\mathrm{w}}_1, \ldots, u^{\mathrm{w}}_R) = \boldsymbol{u}$ to $(X^{\mathrm{s}}_{1,m+1}, \ldots, X^{\mathrm{s}}_{R,m+1}, X^{\mathrm{w}}_{1,m+1}, \ldots, X^{\mathrm{w}}_{R,m+1}) = (v^{\mathrm{s}}_1, \ldots, v^{\mathrm{s}}_R, v^{\mathrm{w}}_1, \ldots, v^{\mathrm{w}}_R) = \boldsymbol{v}$. In this chapter, we define $K_{1,m} = \max\{k : X^{\mathrm{s}}_{k,m} > 0 \text{ for } k = 1, \ldots, R\}$ and $K_{2,m} = \min\{k : X^{\mathrm{w}}_{k,m} > 0 \text{ for } k = 1, \ldots, R\}$ be, respectively, the highest priority index of the customers in service and the priority index of the customer waiting at the first position in queue at time $m\Delta t$ for $m = 0, 1, \cdots$. Let $(Z_{a,m}, Z_{a,m})$ now describe the arrival and departure process during $[m\Delta t, (m+1)\Delta t)$ for $m = 0, 1, \cdots$ where $Z_{a,m}$ monitors the service priority of a customer entering into the queue, $Z_{a,m} \in \{0, \ldots, R\}$. $Z_{a,m}$ is 0 if no customer arrives to the system, and $Z_{d,m}$ monitors the service priority of a customer departing from the system, $Z_{d,m} \in \{j : X^{\mathrm{s}}_{j,m} > 0 \text{ for } j = 1, \ldots, R\}$ or $Z_{d,m} = 0$ if no customer departs from the system.

The way of storing information about the state of the queueing system in a vector over time is analogue to taking snapshots in photography starting at some initial time and then at every time points incremented by a fixed Δt thereafter. By the properties of $M/M/1$ queues, the probability of observing more than one arrival is negligible, the probability of observing more than one departure is negligible, and the probability of observing one arrival and one departure within the same Δt is also negligible.

4.1.1 Steady State Transition Rules

In order to establish the state transition rules, we first specify how customers are arranged under some circumstances in order to properly study the limiting behavior of priority queues. Besides the usual first-come, first-serve rule within each of the priority classes, when there is a customer U_j with priority j in service and a customer from a higher priority class i, $i < j$, enters the queue, we assume the preemptive repeat-different scheduling. It means that if customer U_j in service does not depart from the system, the service of U_j would be interrupted. U_j then would retreat to the first waiting position in queue within the group of class j customers and U_i begins to receive service at the start of the next Δt of time. On the other hand, any customer in service has a certain probability of departing from the system either for completion of treatment or other reasons. We further assume that if a customer arrives during a short interval of time Δt, this customer does not depart from the system within the same Δt.

Under such discipline, when a customer with a service priority level i is in service, it means that there is no customer with service priority score lower than i waiting in the queue. When a preempted lower service priority customer resumes service, the service process starts afresh and the service-time probability distribution is assumed to be the same as before. For fixed values of R, b and c, we establish the state transition rules of an imbedded $M/M/1$-$R/b/c$ preemptive repeat-different priority queue in the following.

When $0 \leq \sum_{i=1}^{R} u_i^{\mathrm{s}} < c$, then for $\ell = 1, \ldots, R$,

$$\begin{aligned} v_\ell^{\mathrm{s}} = &\ u_\ell^{\mathrm{s}} + I(Z_{a,m}, \ell) - I(Z_{d,m}, \ell), \\ v_\ell^{\mathrm{w}} = &\ u_\ell^{\mathrm{w}} = 0, \end{aligned}$$

where

$$I(Z_{a,m}, \ell) = \begin{cases} 1 & \text{if } 0 < Z_{a,m} \leq R \text{ and } \ell = Z_{a,m}, \\ 0 & \text{otherwise}, \end{cases}$$

$$I(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } \sum_{i=1}^{R} u_i^{\mathrm{s}} > 0,\ Z_{d,m} > 0 \text{ and } \ell = Z_{d,m}, \\ 0 & \text{otherwise}, \end{cases}$$

are indicator functions to control that the ℓth component of $(X_{1,m}^{\mathrm{s}}, \ldots, X_{R,m}^{\mathrm{s}})$ should be added by 1, be subtracted by 1, or stays unchanged depending on $Z_{a,m}$ and $Z_{d,m}$. Note that, in our service rule, if $\sum_{i=1}^{R} u_i^{\mathrm{s}} < c$, then $\sum_{i=1}^{R} u_i^{\mathrm{w}} = 0$, also that if $(X_{1,m}^{\mathrm{s}}, \ldots, X_{R,m}^{\mathrm{s}}, X_{1,m}^{\mathrm{w}}, \ldots, X_{R,m}^{\mathrm{w}}) = \mathbf{0}$, then $Z_{d,m} = 0$.

When the threshold on the number of customers in service is reached but not the queue capacity, i.e., $C_m = c$ and $B_m < b$, then

$$\begin{aligned} v_\ell^{\mathrm{s}} &= u_\ell^{\mathrm{s}} + I_1(Z_{a,m}, Z_{d,m}, \ell) - I_2(Z_{a,m}, Z_{d,m}, \ell), \\ v_\ell^{\mathrm{w}} &= u_\ell^{\mathrm{w}} + I_3(Z_{a,m}, Z_{d,m}, \ell) - I_4(Z_{a,m}, Z_{d,m}, \ell), \end{aligned}$$

where $I_k(Z_{a,m}, Z_{d,m}, \ell)$ for $k = 1, \ldots, 4$ are indicator functions determined by the following:

$$I_1(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (C1)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (C2)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (C3)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_4(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (C4)} \\ 0 & \text{otherwise,} \end{cases}$$

where

(C1) $Z_{d,m} = 0$, $0 < Z_{a,m} < K_{1,m}$ and $\ell = Z_{a,m}$, or

$Z_{d,m} > 0$, $Z_{a,m} = 0$, $\sum\limits_{i=1}^{R} u_i^{\mathrm{w}} > 0$ and $\ell = K_{2,m}$, or

$Z_{d,m} > 0$, $Z_{a,m} > 0$, $\sum\limits_{i=1}^{R} u_k^{\mathrm{w}} > 0$ and $\ell = \min(Z_{a,m}, K_{2,m})$, or

$Z_{d,m} > 0$, $Z_{a,m} > 0$, $\sum\limits_{i=1}^{R} u_i^{\mathrm{w}} = 0$ and $\ell = Z_{a,m}$,

(C2) $Z_{d,m} = 0$, $0 < Z_{a,m} < K_{1,m}$ and $\ell = K_{1,m}$, or

$Z_{d,m} > 0$ and $\ell = Z_{d,m}$,

(C3) $Z_{d,m} = 0$, $Z_{a,m} > 0$, $\ell = \max(Z_{a,m}, K_{1,m})$, or

$Z_{d,m} > 0$, $Z_{a,m} \geq K_{2,m} > 0$ and $\ell = Z_{a,m}$,

(C4) $Z_{d,m} > 0$, $Z_{a,m} = 0$ or $Z_{a,m} \geq K_{2,m}$ and $\ell = K_{2,m}$.

Else, when the queue capacity is reached, $B_m = b$, then no new arrival will be accepted to enter the queueing system, $Z_{a,m} = 0$ with probability 1, and

$$\begin{aligned} v_\ell^{\rm s} = & \ u_\ell^{\rm s} + I_1(Z_{d,m}, \ell) - I_2(Z_{d,m}, \ell), \\ v_\ell^{\rm s} = & \ u_\ell^{\rm s} - I_3(Z_{d,m}, \ell). \end{aligned}$$

where

$$I_1(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} > 0 \text{ and } \ell = K_{2,m}, \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} > 0 \text{ and } \ell = Z_{d,m}, \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} > 0 \text{ and } \ell = K_{2,m}, \\ 0 & \text{otherwise.} \end{cases}$$

Example 4.1.1. For an emergency department having parameters $R = 3$, $b = 10$ and $c = 3$, arrivals are categorized as service priority I (highest), II or III (lowest), a working emergency service team allows up to 3 patients at the same time to undergo treatment process, and the department can allow up to 10 patients in the system (ie. allowing up to 7 in the wait room). We give a sample of possible state of arrival as a priority II patient arrives to the ED and a possible state transition with description.

One-step Transition

(0,0,0,0,0,0) → **(0,1,0,0,0,0)** The queue system is empty and treatment begins for any new arrival.

(1,0,1,0,0,0) → (1,1,1,0,0,0) One of each of priority I and III patients are in treatment and are not departing, no one is waiting in queue and the service threshold is not reached, hence treatment begins for a newly arrived priority II patient.

(1,2,0,0,1,1) → (1,2,0,1,2,1) One priority I and two priority II patients are in treatment process and are not departing, hence the new arrival waits in queue behind all other priority II patients already waiting in line.

(1,1,1,0,0,1) → (0,2,1,0,0,1) One patient of each category are undergoing treatment, the priority I patient in treatment is departing while no other priority I or II patient is waiting in line, hence the new arrival begins treatment process at time.

(1,1,1,0,0,1) → (1,2,0,0,0,2) One patient of each priority class are undergoing treatment and no one is departing, the treatment process of the priority III patient is interrupted and is queued to the first waiting position of all patients of priority class III, treatment begins for the newly arrived priority II patient.

□

4.1.2 Transition Probabilities for Finding the Ergodic Distribution

Once the state transition rules with their corresponding probabilities are defined for all states in Ω_X, steady-state probability distribution $\boldsymbol{\pi}$ of all states in the finite state space can be proved to always exist under our assumptions and be obtained first.

Unlike the case of the preemptive repeat-different priority queue capable of providing service to at most one customer, the structure of the transition probabilities now is rather complicated. It is not impossible to lay out the state transition probabilities but rather arduous to do so. However, given $(X^{\mathrm{s}}_{1,m}, \ldots, X^{\mathrm{s}}_{R,m}, X^{\mathrm{w}}_{1,m}, \ldots, X^{\mathrm{w}}_{R,m}) = \boldsymbol{u}$, the one-step transition probabilities $p_{\boldsymbol{uv}} = \Pr\{(X^{\mathrm{s}}_{1,m+1}, \ldots, X^{\mathrm{s}}_{R,m+1}, X^{\mathrm{w}}_{1,m+1}, \ldots, X^{\mathrm{w}}_{R,m+1}) = \boldsymbol{v} \mid (X^{\mathrm{s}}_{1,m}, \ldots, X^{\mathrm{s}}_{R,m}, X^{\mathrm{w}}_{1,m}, \ldots, X^{\mathrm{w}}_{R,m}) = \boldsymbol{u}\}$ now not only depend on a single customer in treatment process, but on the number and the service priority of customers who are undergoing treatment processes by

Proposition 4.0.2. *Consider an experiment in which a certain event occurs with probability $\mu\Delta t + o(\Delta t)$ and does not occur with probability $1 - \mu\Delta t + o(\Delta t)$ as $\Delta t \to 0$, where Δt is arbitrarily small and $0 < \Delta t < \frac{1}{\mu}$. In a number $x > 0$ independent trials of the experiment,*

1. *the probability that the event never occurs in x trials is $(1 - x\mu\Delta t + o(\Delta t))$;*
2. *the probability that the event occurs once in x trials is $x\mu\Delta t + o(\Delta t)$;*
3. *the probability that the event occurs more than once in x trials is $o(\Delta t)$.*

Proof. Let Y denote the number of occurrence of the event in x trials, then the probability that the event never occurs in x trials is

$$\begin{aligned} P(Y=0) = &\ \tbinom{x}{0}(\mu\Delta t + o(\Delta t))^0(1-\mu\Delta t + o(\Delta t))^x \\ = &\ (1-\mu\Delta t + o(\Delta t))^x \\ = &\ \sum_{i=0}^{x}(-1)^i\tbinom{x}{i}1^{x-i}(\mu\Delta t + o(\Delta t))^i \\ = &\ 1-\tbinom{x}{1}(\mu\Delta t + o(\Delta t)) + \tbinom{x}{2}(\mu\Delta t + o(\Delta t))^2 - \cdots \\ = &\ 1 - x\mu\Delta t + o(\Delta t). \end{aligned}$$

By the same token, the probability that the event occurs once in x trials is

$$\begin{aligned} P(Y=1) = &\ \tbinom{x}{1}(\mu\Delta t + o(\Delta t))^1(1-\mu\Delta t + o(\Delta t))^{x-1} \\ = &\ x\cdot(\mu\Delta t + o(\Delta t))\cdot\sum_{i=0}^{x-1}(-1)^i\tbinom{x-1}{i}1^{x-1-i}(\mu\Delta t + o(\Delta t))^i \\ = &\ x\cdot(\mu\Delta t + o(\Delta t))\cdot\big(1-\tbinom{x-1}{1}(\mu\Delta t + o(\Delta t)) + \tbinom{x-1}{2}(\mu\Delta t + o(\Delta t))^2 \\ &\ -\cdots\big) \\ = &\ x\mu\Delta t + o(\Delta t), \end{aligned}$$

and the probability that the event occurs more than once in x trials is

$$\begin{aligned} P(Y\geq 2) = &\ \sum_{y=2}^{x}\tbinom{x}{y}(\mu\Delta t + o(\Delta t))^y(1-\mu\Delta t + o(\Delta t))^{x-y} \\ = &\ \tbinom{x}{2}(\mu\Delta t + o(\Delta t))^2(1-\mu\Delta t + o(\Delta t))^{x-2} + \cdots \\ &\ +\tbinom{x}{x}(\mu\Delta t + o(\Delta t))^x(1-\mu\Delta t + o(\Delta t))^0 \\ = &\ o(\Delta t) \end{aligned}$$

where $o(\Delta t)$ denotes a function of Δt such that $\lim_{\Delta t\to 0} o(\Delta t)/\Delta t = 0$ □

So, in conjunction to the state transition rules, the one-step steady state transition probability $p_{\boldsymbol{uv}}$'s can be calculated according to the following.

If $(X^{\mathrm{s}}_{1,m},\ldots,X^{\mathrm{s}}_{R,m},X^{\mathrm{w}}_{1,m},\ldots,X^{\mathrm{w}}_{R,m}) = \mathbf{0}$, then $Z_{d,m}=0$ and

$$p_{\boldsymbol{uv}} = \begin{cases} \left(1-\sum\limits_{i=1}^{R}\lambda_i\Delta t + o(\Delta t)\right) & \text{if } Z_{a,m}=0, \\ \lambda_{Z_{a,m}}\Delta t + o(\Delta t) & \text{if } Z_{a,m}>0, \text{ for } Z_{a,m}=1,\ldots,R, \\ 0 & \text{otherwise.} \end{cases}$$

If $0 < B_m < b$, we have

$$p_{\boldsymbol{uv}} = \begin{cases} \left(1-\sum\limits_{i=1}^{R}\lambda_i\Delta t + o(\Delta t)\right)\left(1-\sum\limits_{i=1}^{R}u^{\mathrm{s}}_i\mu_i\Delta t + o(\Delta t)\right) & \text{if (1f)} \\ \left(1-\sum\limits_{i=1}^{r}\lambda_i\Delta t + o(\Delta t)\right)\left(u^{\mathrm{s}}_{Z_{d,m}}\mu_{Z_{d,m}}\Delta t + o(\Delta t)\right) & \text{if (2f)} \\ \left(\lambda_{Z_{a,m}}\Delta t + o(\Delta t)\right)\left(1-\sum\limits_{i=1}^{R}u^{\mathrm{s}}_i\mu_i\Delta t + o(\Delta t)\right) & \text{if (3f)} \\ \left(\lambda_{Z_{a,m}}\Delta t + o(\Delta t)\right)\left(u^{\mathrm{s}}_{Z_{d,m}}\mu_{Z_{d,m}}\Delta t + o(\Delta t)\right) & \text{if (4f)} \\ 0 & \text{otherwise,} \end{cases}$$

where

(1f) $Z_{d,m}=0$ and $Z_{a,m}=0$,

(2f) $Z_{d,m}>0$ and $Z_{a,m}=0$

(3f) $Z_{d,m}=0$ and $Z_{a,m}>0$, for $Z_{a,m}=1,\ldots,R$,

(4f) $Z_{d,m}>0$ and $Z_{a,m}>0$, for $Z_{a,m}=1,\ldots,R$.

If $B_m=b$, then $Z_{a,m}=0$ and

$$p_{\boldsymbol{uv}} = \begin{cases} \left(1-\sum\limits_{i=1}^{R}u^{\mathrm{s}}_i\mu_i\Delta t + o(\Delta t)\right) & \text{if } Z_{d,m}=0, \\ u^{\mathrm{s}}_{Z_{d,m}}\mu_{Z_{d,m}}\Delta t + o(\Delta t) & \text{if } Z_{d,m}>0, \\ 0 & \text{otherwise.} \end{cases}$$

4.1.3 Obtaining the Distribution of *Wait Time*

Now we focus on computing the distribution of the length of wait time, $W(\boldsymbol{\alpha}_r)$, of a particular priority r $(1 \leq r \leq R)$ customer who arrived to the queue in the interval Δt just before some t_0 and begins to receive service for the first time later at some time $t_{\alpha_r} \geq t_0$. Our main focus here is to obtain the distribution of the length of time $W(\boldsymbol{\alpha}_r) = t_{\alpha_r} - t_0$.

As a priority r customer is entering the system, the initial state can only be in one of the three categories: (1) the system capacity is full at time $t_0 - \Delta t$ and no new arrival can join the queue; (2) either the service threshold is not reached at time $t_0 - \Delta t$, or there are at least c customers in service during $[t_0 - \Delta t, t_0)$ yet the system capacity is not full and there is at least one customer of class j, $j > r$, in service, therefore, service of the customer who has the highest priority index is interrupted to begin service on the arrived priority r customer starting at time t_0. In this case, the absorbing state $\boldsymbol{\alpha}_r$ is reached in terms of the imbedded Markov chain; or (3) the priority r customer has to wait in queue, indicating that at time t_0, there is no customer with priority index $j > r$ in service.

This section differs from Section §3.2.2 in two aspects. First, as mentioned in the preceding, states of the imbedded Markov chain depend on the number of and the priority class of customers who are in service. Secondly, the state of interest is the state at the instance when the priority r customer of our focus starts to receive service for the first time, not the state at the time of customer's departure. In the modeling of hospital emergency service, the random variable **wait time** can be defined to be the length from the time of customer's triage or registration time to the time of the customer first visited by an emergency physician.

When studying wait time, we use $Y_{rm} = (Y^{\mathrm{s}}_{1,m}, \ldots, Y^{\mathrm{s}}_{r,m}, Y^{\mathrm{w}}_{1,m}, \ldots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m})$ to describe the status of the system at time $m\Delta t$ after t_0 for $m = 0, 1, 2, \cdots$ where $Y^{\mathrm{s}}_{i,m}$ monitors the number of class i customers in service, $i = 1, \ldots, r$, $Y^{\mathrm{w}}_{i,m}$ monitors the number of class i waiting customers, $i = 1, \ldots, r$, $Y^{\mathrm{w}}_{r+1,m}$ records the number of class $j > r$ waiting customers when $m = 0$ and $Y^{\mathrm{w}}_{r+1,m}$ monitors the number of class $j > r$ waiting customers plus the number of class r waiting customers who joined the queue after time t_0 for $m = 1, 2, \cdots$. Y_{rm} induces a finite state space Ω_{Y_r} which consists of states satisfying the following criteria:

$$
\begin{aligned}
\Omega_{Y_r} = \ & \big\{\boldsymbol{\alpha}_r\big\} \bigcup \big\{(y^{\mathrm{s}}_1, \ldots, y^{\mathrm{s}}_r, y^{\mathrm{w}}_1, \ldots, y^{\mathrm{w}}_r, y^{\mathrm{w}}_{r+1}) \mid 0 \le y^{\mathrm{s}}_1 \le c, \\
& \qquad 0 \le y^{\mathrm{w}}_1 \le b - c - 1,\ 0 \le y^{\mathrm{s}}_i \le c - \sum_{k=1}^{i-1} y^{\mathrm{s}}_k \text{ for } i = 2, \ldots, r, \\
& \qquad 0 \le y^{\mathrm{w}}_2 \le b - c - 1 - y^{\mathrm{w}}_1,\ 0 \le y^{\mathrm{w}}_j \le b - c - 1 - \sum_{k=1}^{j-1} y^{\mathrm{w}}_k \\
& \qquad \text{for } j = 3, \ldots, r-1,\ 1 \le y^{\mathrm{w}}_r \le b - c - 1 - \sum_{k=1}^{r-1} y^{\mathrm{w}}_k, \\
& \qquad 0 \le y^{\mathrm{w}}_{r+1} \le b - c - 1 - \sum_{k=1}^{r} y^{\mathrm{w}}_k \text{ and for each} \\
& \qquad (y^{\mathrm{s}}_1, \ldots, y^{\mathrm{s}}_r, y^{\mathrm{w}}_1, \ldots, y^{\mathrm{w}}_r, y^{\mathrm{w}}_{r+1}),\ y^{\mathrm{w}}_k = 0 \text{ for all} \\
& \qquad k = 1, \ldots, k^* - 1\big\}
\end{aligned}
$$

where $k^* = \max\{k : y^{\mathrm{s}}_k > 0 \text{ for } k = 1, \ldots, r \mid (y^{\mathrm{s}}_1, \ldots, y^{\mathrm{s}}_r, y^{\mathrm{w}}_1, \ldots, y^{\mathrm{w}}_r, y^{\mathrm{w}}_{r+1}) \neq \mathbf{0}\}$ in the above. $\boldsymbol{\alpha}_r$ denotes the state of system at the instance t_{α_r}.

For the convenience when describing a one-step state transition and transition probability denoted by $p_{\boldsymbol{uv}} = \Pr\{Y_{[r]m+1} = \boldsymbol{v} \mid Y_{rm} = \boldsymbol{u}\}$ for $m = 0, 1, \cdots$, we use the notation $\boldsymbol{u} \to \boldsymbol{v}$ to describe a one-step state transition of the process going from $(Y^{\text{s}}_{1,m}, \ldots, Y^{\text{s}}_{r,m}, Y^{\text{w}}_{1,m}, \ldots, Y^{\text{w}}_{r,m}, Y^{\text{w}}_{r+1,m}) = (u^{\text{s}}_1, \ldots, u^{\text{s}}_r, u^{\text{w}}_1, \ldots, u^{\text{w}}_r, u^{\text{w}}_{r+1}) = \boldsymbol{u}$ to $(Y^{\text{s}}_{1,m+1}, \ldots, Y^{\text{s}}_{r,m+1}, Y^{\text{w}}_{1,m+1}, \ldots, Y^{\text{w}}_{r,m+1}, Y^{\text{w}}_{r+1,m+1}) = (v^{\text{s}}_1, \ldots, v^{\text{s}}_r, v^{\text{w}}_1, \ldots, v^{\text{w}}_r, v^{\text{w}}_{r+1}) = \boldsymbol{v}$. In this section, we define $K_{3,m} = \max\{k : Y^{\text{s}}_{k,m} > 0 \text{ for } k = 1, \ldots, r\}$ and $K_{4,m} = \min\{k : Y^{\text{w}}_{k,m} > 0 \text{ for } k = 1, \ldots, r\}$ be, respectively, the highest priority index of the customers in service and the priority index of the customer waiting at the first position in queue at time $m\Delta t$ for $m = 0, 1, \cdots$ since t_0. Let $(Z_{a,m}, Z_{a,m})$ now describe the arrival and departure process during $[m\Delta t, (m + 1)\Delta t)$ for $m = 0, 1, \cdots$ since t_0 where $Z_{a,m}$ monitors the service priority of a customer entering into the queue, $Z_{a,m} \in \{0, \ldots, R\}$. $Z_{a,m}$ is 0 if no customer arrives to the system, and $Z_{d,m}$ monitors the service priority of a customer departing from the system, $Z_{d,m} \in \{j : Y^{\text{s}}_{j,m} > 0 \text{ for } j = 1, \ldots, r\}$ or $Z_{d,m} = 0$ if no customer departs from the system.

Suppose a priority r customer arrives and joins the system is essentially in condition (3) above at time t_0, then $Y^{\text{w}}_{r,m} \geq 1$. We will set forth the state transition rules and probabilities for obtaining the conditional, in the sense that the arrived customer is able to join the queue, waiting time distribution. The only case where transition rules and probabilities need to be discussed about is the case when the arrived priority r customer is queued at the position behind all waiting customers of priority score of i or lower.

Transition Rules and Transition Probabilities

When $C_m = c$, $B_m < b$ and $(Y^{\mathrm{s}}_{1,m}, \ldots, Y^{\mathrm{s}}_{r,m}, Y^{\mathrm{w}}_{1,m}, \ldots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m}) = (u^{\mathrm{s}}_1, \ldots, u^{\mathrm{s}}_r, 0, \ldots, 0, u^{\mathrm{w}}_r = 1, u^{\mathrm{w}}_{r+1})$, then

$$
\begin{aligned}
v^{\mathrm{s}}_\ell &= \begin{cases} u^{\mathrm{s}}_\ell + I_1(Z_{a,m}, Z_{d,m}, \ell) - I_2(Z_{a,m}, Z_{d,m}, \ell) \\ \alpha^{\mathrm{s}}_\ell & \text{if } (*) \end{cases} \\
v^{\mathrm{s}}_r &= \begin{cases} u^{\mathrm{s}}_r - I_3(Z_{a,m}, Z_{d,m}) \\ \alpha^{\mathrm{s}}_r & \text{if } (*) \end{cases} \\
v^{\mathrm{w}}_\ell &= \begin{cases} u^{\mathrm{w}}_\ell + I_4(Z_{a,m}, Z_{d,m}, \ell) \\ \alpha^{\mathrm{w}}_\ell & \text{if } (*) \end{cases} \\
v^{\mathrm{w}}_r &= \begin{cases} u^{\mathrm{w}}_r + I_5(Z_{a,m}, Z_{d,m}, \ell) \\ \alpha^{\mathrm{w}}_r & \text{if } (*) \end{cases} \\
v^{\mathrm{w}}_{r+1} &= \begin{cases} u^{\mathrm{w}}_{r+1} + I_6(Z_{a,m}, Z_{d,m}) \\ \alpha^{\mathrm{w}}_{r+1} & \text{if } (*) \end{cases}
\end{aligned}
\tag{4.1.1}
$$

for $\ell = 1, \ldots, r-1$ where $(*)$ in (4.1.1) is the condition: $Z_{d,m} > 0$ and, $Z_{a,m} = 0$ or $Z_{a,m} \geq r$. I_k in (4.1.1) for $k = 1, \ldots, 6$ are defined as in the following:

$$
I_1(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (D1)} \\ 0 & \text{otherwise,} \end{cases}
$$

$$
I_2(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (D2)} \\ 0 & \text{otherwise,} \end{cases}
$$

$$I_3(Z_{a,m}, Z_{d,m}) = \begin{cases} 1 & \text{if (D3)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_4(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (D4)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_5(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (D5)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_6(Z_{a,m}) = \begin{cases} 1 & \text{if (D6)} \\ 0 & \text{otherwise,} \end{cases}$$

where

(D1) $0 < Z_{a,m} < K_{3,m}$, $Z_{d,m} = 0$ and $\ell = Z_{a,m}$, or if $0 < Z_{a,m} < r$, $Z_{d,m} > 0$ and $\ell = Z_{a,m}$,

(D2) $0 < Z_{a,m} < K_{3,m}$, $Z_{d,m} = 0$ and $\ell = K_{3,m}$, or if $0 < Z_{a,m} < r$ and $\ell = Z_{d,m} > 0$,

(D3) $u_r^{\mathrm{s}} > 0$, $0 < Z_{a,m} < r$, and $Z_{d,m} = 0$ or $Z_{d,m} = r$,

(D4) $Z_{d,m} = 0$, $0 < Z_{a,m} < K_{3,m}$ and $\ell = K_{3,m}$, or if $Z_{d,m} > 0$, $K_{3,m} < \ell = Z_{a,m} < r$,

(D5) $u_r^{\mathrm{s}} > 0$, $0 < Z_{a,m} < r$, and $Z_{d,m} = 0$,

(D6) $Z_{a,m} \geq r$ and $Z_{d,m} = 0$.

If $C_m = c$, $B_m < b$, $1 < \sum_{k=1}^{r} u_k^{\mathrm{s}}$ and $u_r^{\mathrm{w}} \geq 1$, then

$$\begin{aligned}
v_\ell^{\mathrm{s}} = &\ u_\ell^{\mathrm{s}} + I_1(Z_{a,m}, Z_{d,m}, \ell) - I_2(Z_{a,m}, Z_{d,m}, \ell), \\
v_r^{\mathrm{s}} = &\ u_r^{\mathrm{s}} + I_3(Z_{a,m}, Z_{d,m}) - I_4(Z_{a,m}, Z_{d,m}), \\
v_\ell^{\mathrm{w}} = &\ u_\ell^{\mathrm{w}} + I_5(Z_{a,m}, Z_{d,m}, \ell) - I_6(Z_{a,m}, Z_{d,m}, \ell), \\
v_r^{\mathrm{w}} = &\ u_r^{\mathrm{w}} + I_7(Z_{a,m}, Z_{d,m}) - I_8(Z_{a,m}, Z_{d,m}), \\
v_{r+1}^{\mathrm{w}} = &\ u_{r+1}^{\mathrm{w}} + I_9(Z_{a,m})
\end{aligned}$$

for $\ell = 1, \ldots, r-1$, where

$$I_1(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (E1)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (E2)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(Z_{a,m}, Z_{d,m}) = \begin{cases} 1 & \text{if (E3)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_4(Z_{a,m}, Z_{d,m}) = \begin{cases} 1 & \text{if (E4)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_5(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (E5)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_6(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (E6)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_7(Z_{a,m}, Z_{d,m}) = \begin{cases} 1 & \text{if (E7)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_8(Z_{a,m}, Z_{d,m}) = \begin{cases} 1 & \text{if (E8)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_9(Z_{a,m}) = \begin{cases} 1 & \text{if (E9)} \\ 0 & \text{otherwise,} \end{cases}$$

where

(E1) $Z_{d,m} = 0$, $0 < Z_{a,m} < K_{3,m}$ and $\ell = Z_{a,m}$, or if $Z_{d,m} > 0$, $0 < Z_{a,m} < r$ and $\ell = \min(Z_{a,m}, K_{4,m})$, or if $Z_{d,m} > 0$, $Z_{a,m} = 0$, and $\ell = K_{4,m}$,

(E2) $Z_{d,m} = 0$, $0 < Z_{a,m} < K_{3,m}$ and $\ell = K_{3,m}$, or if $Z_{d,m} > 0$ and $\ell = Z_{d,m}$,

(E3) $\sum_{k=1}^{r-1} u_k^{\mathrm{w}} = 0$, $u_r^{\mathrm{w}} > 1$, $Z_{d,m} > 0$, and $Z_{a,m} = 0$ or $Z_{a,m} \geq r$,

(E4) $\sum_{k=1}^{r-1} u_k^{\mathrm{w}} = 0$, $u_r^{\mathrm{s}} > 0$, $Z_{d,m} = 0$ and $Z_{a,m} < r$, or if $\sum_{k=1}^{r-1} u_k^{\mathrm{w}} = 0$, $u_r^{\mathrm{s}} > 0$ and $Z_{d,m} = r$,

(E5) $Z_{d,m} = 0$, $0 < Z_{a,m} < r$ and $\ell = \max(Z_{a,m}, K_{3,m})$, or if $Z_{d,m} > 0$, $K_{r,m} \leq Z_{a,m} < r$ and $\ell = Z_{a,m}$,

(E6) $Z_{d,m} > 0$, $Z_{a,m} = 0$ or $Z_{a,m} \geq K_{4,m}$, and $\ell = K_{4,m}$,

(E7) $Z_{d,m} = 0$, $0 < Z_{a,m} < r$ and $u_r^{\mathrm{s}} > 0$,

(E8) $\sum_{k=1}^{r-1} u_k^{\mathrm{w}} = 0$, $Z_{d,m} > 0$ and, $Z_{a,m} = 0$ or $Z_{a,m} \geq r$, (E9) $Z_{a,m} \geq r$.

When $B_m = b$, $u_k^{\mathrm{w}} = 0$ for $k = 1, \ldots, r-1$ and $u_r^{\mathrm{w}} = 1$, then $Z_{a,m} = 0$ and

$$Y_{[r](m+1)} = \begin{cases} \boldsymbol{\alpha}_r & \text{if } Z_{d,m} > 0, \\ Y_{rm} & \text{otherwise,} \end{cases}$$

else, if $u_k^{\mathrm{w}} > 0$ for some $k = 1, \ldots, r-1$ or $u_r^{\mathrm{w}} > 1$, then $Z_{a,m} = 0$ and for $\ell = 1, \ldots, r$,

$$\begin{aligned} v_\ell^{\mathrm{s}} = &\ u_\ell^{\mathrm{s}} + I_1(Z_{d,m}, \ell) - I_2(Z_{d,m}, \ell) \\ v_\ell^{\mathrm{w}} = &\ u_\ell^{\mathrm{w}} - I_3(Z_{d,m}, \ell) \\ v_{r+1}^{\mathrm{w}} = &\ u_{r+1}^{\mathrm{w}} \end{aligned}$$

where

$$I_1(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} > 0 \text{ and } \ell = K_{4,m}, \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} > 0 \text{ and } \ell = Z_{d,m}, \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if } Z_{d,m} > 0 \text{ and } \ell = K_{4,m}, \\ 0 & \text{otherwise.} \end{cases}$$

In conjunction to the state transition rules above, the one-step transition probability matrix used to obtain the waiting-time distribution of $\boldsymbol{\alpha}_r$ can be constructed according to the the following conditions.

If $C_m = c$ and $B_m < b$,

$$p_{\boldsymbol{uv}} = \begin{cases} \left(1 - \sum\limits_{k=1}^{R} \lambda_k \Delta t + o(\Delta t)\right)\left(1 - \sum\limits_{k=1}^{r} u_k^{\mathrm{s}} \mu_k \Delta t + o(\Delta t)\right) & \text{if (1g)} \\ \left(1 - \sum\limits_{k=1}^{R} \lambda_k \Delta t + o(\Delta t)\right)(u_{Z_{d,m}}^{\mathrm{s}} \mu_{Z_{d,m}} \Delta t + o(\Delta t)) & \text{if (2g)} \\ (\lambda_{Z_{a,m}} \Delta t + o(\Delta t))\left(1 - \sum\limits_{k=1}^{r} u_k^{\mathrm{s}} \mu_k \Delta t + o(\Delta t)\right) & \text{if (3g)} \\ (\lambda_{Z_{a,m}} \Delta t + o(\Delta t))(u_{Z_{d,m}}^{\mathrm{s}} \mu_{Z_{d,m}} \Delta t + o(\Delta t)) & \text{if (4g)} \\ 0 & \text{otherwise,} \end{cases}$$

where

(1g) $Z_{d,m} = 0$ and $Z_{a,m} = 0$,

(2g) $Z_{d,m} > 0$ and $Z_{a,m} = 0$,

(3g) $Z_{d,m} = 0$ and $Z_{a,m} > 0$, for $Z_{a,m} = 1, \ldots, R$,

(4g) $Z_{d,m} > 0$ and $Z_{a,m} > 0$, for $Z_{a,m} = 1, \ldots, R$.

If $C_m = c$ and $B_m = b$, then

$$p_{\boldsymbol{uv}} = \begin{cases} \left(1 - \sum\limits_{k=1}^{r} u_k^{\mathrm{s}} \mu_k \Delta t + o(\Delta t)\right) & \text{if } Z_{d,m} = 0 \text{ and } Z_{a,m} = 0, \\ u_{Z_{d,m}}^{\mathrm{s}} \mu_{Z_{d,m}} \Delta t + o(\Delta t) & \text{if } Z_{d,m} > 0 \text{ and } Z_{a,m} = 0, \\ 0, & \text{otherwise.} \end{cases}$$

For any $m \geq 1$, if $Y_{rm} = \boldsymbol{\alpha}_r$, then the one-step transition probability is $p_{\boldsymbol{\alpha}_r \to \boldsymbol{\alpha}_r} = 1$.

Once the transition probability matrix M_r is constructed, the distribution of wait time $P(W(\boldsymbol{\alpha}_r) \leq n \mid (Y_{1,0}^{\mathrm{s}}, \ldots, Y_{r,0}^{\mathrm{s}}, Y_{1,0}^{\mathrm{w}}, \ldots, Y_{r,0}^{\mathrm{w}}, Y_{r+1,0}^{\mathrm{w}}))$ can again easily be calculated by

$$P(W(\boldsymbol{\alpha}_r) \leq n \mid \boldsymbol{\xi}_r) = \boldsymbol{\xi}_r M_r^n \boldsymbol{c}'. \tag{4.1.2}$$

However, remember that (4.1.2) here is a conditional distribution in the sense that we assumed in the first place a priority class r customer was able to join the queue at some time t_0 in order for this distribution to be meaningful.

As a result, a meaningful mean and variance of the variable $W(\boldsymbol{\alpha}_r)$ can again be calculated by

$$E[W(\boldsymbol{\alpha}_r)] = \boldsymbol{\xi}_r (I - N_r)^{-1} \mathbf{1}',$$

and

$$\mathrm{Var}(W(\boldsymbol{\alpha}_r)) = \boldsymbol{\xi}_r (I - N_r)^{-1} (I + N_r - \mathbf{1}'\boldsymbol{\xi}_r)(I - N_r)^{-1} \mathbf{1}',$$

see Section §3.2.3 and §3.2.4.

Example 4.1.2. Here we will demonstrate through an example of computing the waiting time distribution, expected wait time and the standard deviation of wait time of patients in a $M/M/1$-$3/b/c$ preemptive repeat-different priority queueing system. A data set was obtained with authorization from the Changhua Christian Hospital Erlin Branch in Taiwan. Many variables were provided and the following variables (English translated) in the data set were used for analysis:

- Patient Triage and Acuity Scale upon Arrival;
- Patient Registration Time and Date;
- Time of First Physician Order Issued;

- Time and Date of Patient Departure from the ED.

Variables such as Patient Triage Time and Time of Initial Assessment by a Physician were not given, therefore we treat the time from patient registration to the time of first order given by an EP after initial examination as patient wait time. The time from the first order given by an EP after initial examination to the time of patient departure from ED is defined as the length of treatment time in this example. In the Changhua Christian Hospital Erlin Branch patients visiting the ED actually can be of four priority categories. But, since the data set from January 01 to December 31 of 2007 contains only five priority IV patients, we decided to combine the five priority IV patient records with the priority III group to perform analyses. Within the 2007 fiscal year, the following descriptive statistics were calculated and tabulated in Table 4.1. Average wait times and treatment times are measured in minutes. It is surprising to us that the rates of arrival and the average treatment times are quite dissimilar among the three priority classes, yet the average wait times (defined in this context)

Table 4.1: Raw Statistics

Patient Priority	Total Number of Visits to ED	Average Wait Time (Std. Dev.)	Average Treatment Time
I	3,501	7.42 (7.01)	78.5036
II	9,267	7.3644 (6.01)	92.8010
III	13,918	7.9785 (5.61)	103.6516

are quite similar (between 7 to 8 minutes).

If we analyze the wait time as if patients are being seen without priority discipline but only by FCFS rule, then the calculated overall average wait time is 7.692 minutes with standard deviation of 5.957 minutes from the data. If we were to use the FMCI method and assuming first-come, first-serve of a single server queue, the estimated mean wait time and standard deviation is summarized in Table 4.2 with respect to parameter values of $b = 11$, $b = 12$ and $b = 13$, and reasonably assigned $c = 7$.

Table 4.2: Estimated Mean Wait-Time in minutes (Standard Deviation)

	$b = 11$	$b = 12$	$b = 13$
$c = 7$	6.8617 (16.9007)	8.1723 (19.4408)	9.2338 (21.5803)

The estimated mean wait times look very reasonable in this FCFS setting. Withal, no precise parameter values of the Changhua Christian Hospital Erlin Branch Emergency Department were given to us specifically regarding the number of patients an emergency medical service team can attend to and the maximum number of patents allowed in the emergency department.

If we model the patient wait times by the FMCI method and assuming a $M/M/1$-$3/b/c$ preemptive repeat-different priority queue, we summarize the expected wait times with respect to different values of b and c in Table 4.3. In the table, we use 0^+ to denote positive real values that are close to 0. Looking at the estimation results and compare to the results in Tables 4.1 and 4.2, we began to suspect that patients are

Table 4.3: Estimated Mean Wait-Time in Minutes

b	c	I	II	III
10	6	0^+	0.097	19.437
11	6	0^+	0.116	22.775
12	6	0^+	0.130	25.259
13	7	0^+	0.014	5.552
14	7	0^+	0.015	5.691
15	7	0^+	0.015	5.779

not always being seen according to the rules of a preemptive repeat-different priority queue.

Later we consulted with one of the Changhua Christian Hospital Erlin Branch emergency departmental staff and were confirmed that in the emergency department there are seven beds available in total including one for first aid purposes. There is always one physician on duty in an eight-hour shift rotation through out a day. Each regular shift then is reduced to six hours during holidays. For a very rough estimate, an emergency medical service team can attend to 5 or 6 patients at times and post doctor attendance there are nurse practitioner (NP) to assist in writing physician's orders and report on examination and treatment procedures. The emergency department can hold 13 to 15 patients being treated or waiting, and additional patient bed space can be arranged for service when confronted by an unforeseen large volume influx.

We wish to extend this modeling in the future by looking at cases where there is more than one emergency medical service teams available (multi-channel priority queue problem) attending multiple patients and allowing multiple departures, within a Δt, from the hospital. This may more closely resemble the practice in bigger medical centres and therefore requires estimates in its own setting. But in the next section, we would like to demonstrate again the application of the finite Markov chain imbedding technique on the Non-preemptive priority queueing model.

4.2 Non-preemptive Model

In contrast to Section §4.1, we will modify the preemptive-repeat priority queueing model to a non-preemptive one. By non-preemptive, suppose a customer has arrived at the system and is assigned with priority score i ($1 \leq i \leq R$). The priority i customer starts receiving service immediately if the current number of customers in service is less than c or is equal to c with one departure occurring. Otherwise, if the the total number of customers in the system is less than b, the newly arrived would be queued in the waiting line even if there is at least one customer of priority score $j > i$ in service. Services in progress are never interrupted as is possible in the preemptive priority queues. The arrived customer with priority number i begins to receive service only when there is no customer of priority score of $k \leq i$ waiting ahead in queue, a customer departs from the system and there is no arrival of a higher service priority customer before service began. We translate such service discipline into a mathematical form in terms of the state transition rules in two parts for the purpose of: studying the limiting system behavior; and obtaining the waiting time distribution.

4.2.1 Transition Rules and Probabilities for Finding Ergodic Distribution of the System

The imbedding procedure is very similar to the case of the preemptive priority queueing model but requiring small modifications on state transition rules to adapt to the non-preemptive feature. We will see next an example of state transition which differs from those in the preemptive model. For an instance of the $M/M/1$-$R/b/c$ non-preemptive priority queue when $R = 2$, $b = 5$, and $c = 2$, suppose the current state of the system is $(X^{\mathrm{s}}_{1,m}, X^{\mathrm{s}}_{2,m}, X^{\mathrm{w}}_{1,m}, X^{\mathrm{w}}_{2,m}) = (1, 1, 0, 1)$: the state having one non-urgent customer and no urgent customer waiting in queue and one of each urgent and non-urgent customers in service. Suppose there is an arrival of an urgent customer and neither of the customers in service had departed during the Δt of the new arrival, then the next state of the system should be:

$$(X^{\mathrm{s}}_{1,m+1}, X^{\mathrm{s}}_{2,m+1}, X^{\mathrm{w}}_{1,m+1}, X^{\mathrm{w}}_{2,m+1}) = \begin{cases} (2,0,0,2) & \text{under the preemptive priority} \\ & \text{scheme,} \\ (1,1,1,1) & \text{under the non-preemptive priority} \\ & \text{scheme.} \end{cases}$$

Next we establish the general state transition rules under different conditions of the system as we did for the preemptive model.

For the convenience when describing one-step state transition rules and probabilities, let $(X^{\mathrm{s}}_{1,m},\ldots,X^{\mathrm{s}}_{R,m},X^{\mathrm{w}}_{1,m},\ldots,X^{\mathrm{w}}_{R,m}) = \boldsymbol{u}$ and $(X^{\mathrm{s}}_{1,m+1},\ldots,X^{\mathrm{s}}_{R,m+1},X^{\mathrm{w}}_{1,m+1},\ldots,$ $X^{\mathrm{w}}_{R,m+1}) = \boldsymbol{v}$ as we did in the previous. The stochastic process $\{(X^{\mathrm{s}}_{1,m},\ldots,X^{\mathrm{s}}_{R,m},X^{\mathrm{w}}_{1,m},$ $\ldots,X^{\mathrm{w}}_{R,m})$ for $m = 0,1,\cdots\}$ induces a finite state space consisting of states satisfying the following criteria:

$$\begin{aligned}\Omega_X = \; & \big\{\mathbf{0}\big\}\bigcup\big\{X = (x^{\mathrm{s}}_1,\ldots,x^{\mathrm{s}}_R,x^{\mathrm{w}}_1,\ldots,x^{\mathrm{w}}_R) \mid 0 \le x^{\mathrm{s}}_1 \le c, 0 \le x^{\mathrm{w}}_1 \le b-c,\\ & \quad 0 \le x^{\mathrm{s}}_i \le c - \sum_{k=1}^{i-1} x^{\mathrm{s}}_k \text{ for } i = 2,\ldots,R,\\ & \quad 0 \le x^{\mathrm{w}}_i \le b - c - \sum_{k=1}^{i-1} x^{\mathrm{w}}_k \text{ for } i = 2,\ldots,R\big\}\end{aligned}$$

When the system is under the conditions that the threshold on the number of customers can be in service is not reached or the queue capacity of the system is reached, the transition rules are exactly the same as the ones shown in the preemptive-repeat priority queues. In other words, the preemptive or non-preemptive feature does not affect state transitions when the system is in one of the above two conditions. The time when we will see differences is when $C_m = c$ and $B_m < b$, for $\ell = 1,\ldots,R$,

$$\begin{aligned}v^{\mathrm{s}}_{\ell,m} = \; & u^{\mathrm{s}}_{\ell,m} + I_1(Z_{a,m},Z_{d,m},\ell) - I_2(Z_{d,m},\ell)\\ v^{\mathrm{w}}_{\ell,m} = \; & u^{\mathrm{w}}_{\ell,m} + I_3(Z_{a,m},\ell) - I_4(Z_{a,m},Z_{d,m},\ell)\end{aligned}$$

where $I_1(Z_{a,m},Z_{d,m},\ell)$, $I_2(Z_{a,m},Z_{d,m},\ell)$, $I_3(Z_{a,m},\ell)$ and $I_4(Z_{a,m},Z_{d,m},\ell)$ are indicator functions determined by $Z_{a,m}$ and $Z_{d,m}$ as in the following:

$$I_1(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (F1)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (F2)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(Z_{a,m}, \ell) = \begin{cases} 1 & \text{if (F3)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_4(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (F4)} \\ 0 & \text{otherwise,} \end{cases}$$

where

(F1) $Z_{d,m} > 0$, $Z_{a,m} = 0$ or $Z_{a,m} \geq K_{2,m}$, and $\ell = K_{2,m}$, or if $Z_{d,m} > 0$, $0 < Z_{a,m} < K_{2,m}$ and $\ell = Z_{a,m}$, or if $Z_{d,m} > 0$, $Z_{a,m} > 0$, $\sum_{k=1}^{R} u_k^{\mathrm{w}} = 0$ and $\ell = Z_{a,m}$,

(F2) $Z_{d,m} > 0$ and $\ell = Z_{d,m}$,

(F3) $Z_{a,m} > 0$ and $\ell = Z_{a,m}$,

(F4) $Z_{d,m} > 0$, $\sum_{k=1}^{R} u_k^{\mathrm{w}} > 0$, $Z_{a,m} = 0$ or $Z_{a,m} \geq K_{2,m}$ and $\ell = K_{2,m}$, or if $Z_{d,m} > 0$, $\sum_{k=1}^{R} u_k^{\mathrm{w}} = 0$ or $0 < Z_{a,m} < K_{2,m}$ and $\ell = Z_{a,m}$.

Example 4.2.1. For an emergency department having parameters $R = 3$, $b = 10$ and $c = 3$, arrivals are categorized as service priority I (highest), II or III (lowest), a working emergency service team allows up to 3 patients at the same time to undergo treatment process, and the department can allow up to 10 patients in the system (ie. allowing up to 7 in the wait room). We give a sample of possible status of the system as a new patient arrives to the ED, the state transitions given the state of arrival:

$(0, 0, 0, 0, 0, 0) \rightarrow (0, 0, 0, 1, 0, 0)$ The emergency department is empty and treatment begins immediately on a newly arrived third priority patient.

$(0, 0, 0, 1, 0, 1) \rightarrow (0, 0, 0, 1, 1, 1)$ One of each first and third priority patient is in treatment and not departing, there is no one waiting in queue, hence, treatment begins on a newly arrived second priority patient.

$(0, 1, 0, 0, 2, 1) \rightarrow (0, 2, 0, 0, 2, 1)$ A first priority and two second priority patients are in treatment process and are not departing, hence the new entry waits in queue behind all other first, second or third priority patients in the queue.

$(1, 0, 0, 1, 1, 1) \rightarrow (1, 0, 0, 1, 1, 1)$ One patient of each category are undergoing treatment, the second priority patient in treatment is departing while no other first and second priority patient is waiting in queue, hence a newly arrived second priority patients begins treatment process.

$(1, 1, 1, 1, 1, 1) \rightarrow (1, 1, 2, 1, 1, 1)$ One patient of each category are undergoing treatment and no one is departing, a newly arrived first priority patient waits in queue behind other first priority patients but in front of all other second and third priority patients already waiting.

In conjunction to the above new transition rules, given $X_m = \boldsymbol{u}$, the one-step state transition probabilities $p_{\boldsymbol{uv}}$ can be constructed the same way and can be summarized in a similar fashino as in the preemptive repeat-different priority queueing model. Various information can be derived from the ergodic distribution of all states in Ω_X. In the next section, we set forth the state transition rules when the interest is on wait time.

4.2.2 Obtaining the Distribution of *Wait Time*

Now we focus on computing the distribution of the length of wait time, $W(\boldsymbol{\alpha}_r)$, of a particular priority r $(1 \leq r \leq R)$ customer who arrived to the queue in the interval Δt just before some t_0 and begins to receive service for the first time later at some time $t_{\alpha_r} \geq t_0$. Our main focus here is to obtain the distribution of the length of time $W(\boldsymbol{\alpha}_r) = t_{\alpha_r} - t_0$.

When studying wait time, we imbed the system information in a vector $Y_{rm} = (Y^{\mathrm{s}}_{1,m}, \ldots, Y^{\mathrm{s}}_{R,m}, Y^{\mathrm{w}}_{1,m}, \ldots, Y^{\mathrm{w}}_{r,m}, Y^{\mathrm{w}}_{r+1,m})$ to describe the status of the system at time $m\Delta t$ after t_0 for $m = 0, 1, 2, \cdots$ where $Y^{\mathrm{s}}_{i,m}$ monitors the number of class i customers in service, $i = 1, \ldots, R$, $Y^{\mathrm{w}}_{j,m}$ monitors the number of class j waiting customers, $j = 1, \ldots, r$, $Y^{\mathrm{w}}_{r+1,m}$ records the number of class $j > r$ waiting customers when $m = 0$ and $Y^{\mathrm{w}}_{r+1,m}$ monitors the number of class $j > r$ waiting customers plus the number of class r waiting customers who joined the queue after time t_0 for $m = 1, 2, \cdots$. Notice that it requires more amount of system information to obtain the waiting time distribution of a priority queue under the non-preemptive scheduling, and Y_{rm} induces a finite state space Ω_{Y_r} which consists of states satisfying the following criteria

$$\begin{aligned}\Omega_{Y_r} = \quad & \{\boldsymbol{\alpha}_r\} \bigcup \{(y_1^{\text{s}}, \ldots, y_R^{\text{s}}, y_1^{\text{w}}, \ldots, y_r^{\text{w}}, y_{r+1}^{\text{w}}) \mid \; 0 \le y_1^{\text{s}} \le c, \\ & \qquad 0 \le y_1^{\text{w}} \le b-c-1, \; 0 \le y_i^{\text{s}} \le c - \sum_{k=1}^{i-1} y_k^{\text{s}} \text{ for } i = 2, \ldots, R, \\ & \qquad 0 \le y_2^{\text{w}} \le b-c-1-y_1^{\text{w}}, \; 0 \le y_j^{\text{w}} \le b-c-1-\sum_{k=1}^{j-1} y_k^{\text{w}} \\ & \qquad \text{for } j = 3, \ldots, r-1, \; 1 \le y_r^{\text{w}} \le b-c-1-\sum_{k=1}^{r-1} y_k^{\text{w}} \\ & \qquad \text{and } 0 \le y_{r+1}^{\text{w}} \le b-c-1-\sum_{k=1}^{r} y_k^{\text{w}}\}\end{aligned}$$

where $\boldsymbol{\alpha}_r$ denotes the state of system at the instance t_{α_r}.

Differences in the state space Ω_{Y_r} under the preemptive and non-preemptive priority scheduling is obvious, the differences in state transition rules and probabilities can easily be identified as we lay the transition rules out under the non-preemptive scheduling.

Transition Rules and Transition Probabilities

For $m = 1, 2, \cdots$, if $C_m = c$, $B_m < b$ and $(Y_{1,m}^{\text{s}}, \ldots, Y_{R,m}^{\text{s}}, Y_{1,m}^{\text{w}}, \ldots, Y_{r,m}^{\text{w}}, Y_{r+1,m}^{\text{w}}) = (u_1^{\text{s}}, \ldots, u_R^{\text{s}}, 0, \ldots, 0, u_r^{\text{w}} = 1, u_{r+1}^{\text{w}})$, then

$$v_\ell^{\text{s}} = \begin{cases} u_\ell^{\text{s}} + I_1(Z_{a,m}, Z_{d,m}, \ell) - I_2(Z_{a,m}, Z_{d,m}, \ell) & \\ \alpha_\ell^{\text{s}} & \text{if } (*), \end{cases}$$

$$v_j^{\text{w}} = \begin{cases} u_j^{\text{w}} + I_3(Z_{a,m}, Z_{d,m}, j) & \\ \alpha_j^{\text{w}} & \text{if } (*), \end{cases}$$

$$v_r^{\text{w}} = \begin{cases} u_r^{\text{w}} & \text{if } Z_{d,m} = 0, \\ \alpha_r^{\text{w}} & \text{if } (*), \end{cases}$$

$$v_{r+1}^{\text{w}} = \begin{cases} u_{r+1}^{\text{w}} + I_4(Z_{a,m}, Z_{d,m}) & \\ \alpha_{r+1}^{\text{w}} & \text{if } (*), \end{cases}$$

for $\ell = 1, \ldots, R$ and $j = 1, \ldots, r-1$, where the condition $(*)$ in the above is: if $Z_{d,m} > 0$ and, $Z_{a,m} = 0$ or $Z_{a,m} \geq r$. This is when there is any departure and either there is no arrival or there is an arrival and the service priority of the arrived customer must be of r or greater. The indicator functions I_k for $k = 1, \ldots, 4$ in the above are:

$$I_1(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (G1)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (G2)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(Z_{a,m}, Z_{d,m}, j) = \begin{cases} 1 & \text{if (G3)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_4(Z_{a,m}, Z_{d,m}) = \begin{cases} 1 & \text{if (G4)} \\ 0 & \text{otherwise,} \end{cases}$$

where

(G1) $Z_{d,m} > 0$, $0 < Z_{a,m} < r$ and $\ell = Z_{a,m}$,

(G2) $Z_{d,m} > 0$, $0 < Z_{a,m} < r$ and $\ell = Z_{d,m}$,

(G3) $Z_{d,m} = 0$, $0 < Z_{a,m} < r$ and $j = Z_{a,m}$, (G4) $Z_{d,m} = 0$ and $Z_{a,m} \geq r$.

If $C_m = c$, $B_m < b$, $u^{\rm w}_r > 1$ or, $u^{\rm w}_r \geq 1$ and $u^{\rm w}_j > 0$ for some $j \in \{1, \ldots, r-1\}$, then for $m = 0, 1, 2, \cdots$, the state transition rules are

$$\begin{aligned} v^{\rm s}_\ell &= u^{\rm s}_\ell + I_1(Z_{a,m}, Z_{d,m}, \ell) - I_2(Z_{d,m}, \ell) \\ v^{\rm w}_j &= u^{\rm w}_j + I_3(Z_{a,m}, j) - I_4(Z_{a,m}, Z_{d,m}, j) \\ v^{\rm w}_r &= u^{\rm w}_r - I_5(Z_{a,m}, Z_{d,m}) \\ v^{\rm w}_{r+1} &= u^{\rm w}_{r+1} + I_6(Z_{a,m}) \end{aligned}$$

for $\ell = 1, \ldots, R$, and $j = 1, \ldots, r-1$,

$$I_1(Z_{a,m}, Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (H1)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_2(Z_{d,m}, \ell) = \begin{cases} 1 & \text{if (H2)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_3(Z_{a,m}, j) = \begin{cases} 1 & \text{if (H3)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_4(Z_{a,m}, Z_{d,m}, j) = \begin{cases} 1 & \text{if (H4)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_5(Z_{a,m}, Z_{d,m}) = \begin{cases} 1 & \text{if (H5)} \\ 0 & \text{otherwise,} \end{cases}$$

$$I_6(Z_{a,m}) = \begin{cases} 1 & \text{if (H6)} \\ 0 & \text{otherwise.} \end{cases}$$

(H1) $Z_{d,m} > 0$, $0 < Z_{a,m} < r$ and $\ell = \min(Z_{a,m}, K_{4,m})$, or if $Z_{d,m} > 0$, $Z_{a,m} = 0$ and $\ell = K_{4,m}$,

(H2) $Z_{d,m} > 0$ and $\ell = Z_{d,m}$,

(H3) $0 < Z_{a,m} < r$ and $j = Z_{a,m}$,

(H4) $Z_{d,m} > 0$, $Z_{a,m} = 0$ and $j = K_{4,m}$, or if $Z_{d,m} > 0$, $0 < Z_{a,m} < r$ and $j = \min(Z_{a,m}, K_{4,m})$,

(H5) $u^{\mathrm{w}}_k = 0$ for $k = 1, \ldots, r-1$, $u^{\mathrm{w}}_r > 1$, $Z_{d,m} > 0$, $Z_{a,m} = 0$ or $Z_{a,m} \geq r$,

(H6) $Z_{a,m} \geq r$

where the variable $K_{4,m}$ is defined as in Section §4.1.3.

If $B_m = b$ and: (1) if $u^{\mathrm{w}}_k = 0$ for $k = 1, \ldots, r-1$ and $u^{\mathrm{w}}_r = 1$, then $Z_{a,m} = 0$ and

$$\boldsymbol{v} = \begin{cases} \boldsymbol{\alpha}_r & \text{if } Z_{d,m} > 0, \\ \boldsymbol{u} & \text{otherwise;} \end{cases}$$

(2) if $u^{\mathrm{w}}_k > 0$ for some $k \in \{1, \ldots, r-1\}$ or $u^{\mathrm{w}}_r > 1$, then $Z_{a,m} = 0$ and for $\ell = 1, \ldots,$ R and $j = 1, \ldots, r$,

$$\begin{aligned} v^{\mathrm{s}}_\ell = &\ u^{\mathrm{s}}_\ell + I_1(Z_{d,m}, \ell) - I_2(Z_{d,m}, \ell) \\ v^{\mathrm{w}}_j = &\ u^{\mathrm{w}}_j - I_3(Z_{d,m}, j) \\ v^{\mathrm{w}}_{r+1} = &\ u^{\mathrm{w}}_{r+1} \end{aligned}$$

where

$$I_1(Z_{d,m}, \ell) = \begin{cases} 1, & \text{if } Z_{d,m} > 0 \text{ and } \ell = K_{4,m}, \\ 0, & \text{otherwise,} \end{cases}$$

$$I_2(Z_{d,m}, \ell) = \begin{cases} 1, & \text{if } Z_{d,m} > 0 \text{ and } \ell = Z_{d,m}, \\ 0, & \text{otherwise,} \end{cases}$$

$$I_3(Z_{d,m}, j) = \begin{cases} 1, & \text{if } Z_{d,m} > 0 \text{ and } j = K_{4,m}, \\ 0, & \text{otherwise.} \end{cases}$$

By the above rules, state transition probabilities $p_{\boldsymbol{uv}}$ can easily be determined depending on $Z_{a,m}$ and $Z_{d,m}$ as for the preemptive priority queues. Once the transition probability matrix is constructed and partitioned as in (3.2.16), the conditional waiting time distribution and moments of the variable wait time can all be obtained for a non-preemptive priority queue.

In this chapter, we clearly see that the finite Markov chain imbedding technique provides us an unified approach to study priority queues. Not only can we know the limiting behavior of a system by examining the steady states, it also conveniently provides a method to obtain the distribution of *wait time* whose definition is custom from problem to problem and from investigator to investigator. We have adopted the methodology found in [10] and formulate the wait time problem into an analysis of first passage time of a particular *pattern* in an induced finite state space of a Markov chain.

Chapter 5

Discussion - Misclassification Effect on Priority Queue

Up till now we have been assuming that there is no error in the assignment of the relative priority of treatment (service) to patients (customers). But in application, such assumption may not be practical. Taking a close look at ED visits, arrivals first enter the triage station and be assessed by a nurse for signs and symptoms. It is expected that the triage process should not take up much time. In emergency medicine there are growing number of reviews and studies on the validity of triage systems used in different regions. To name a few of the triage systems, there are the Australasian Triage Scale (ATS), the Canadian Triage Acuity Scale (CTAS), the Paediatric Canadian Triage and Acuity Score (paedCTAS), the Manchester Triage Scale (MTS), the Emergency Severity Index (ESI) and the Singapore Patient Acuity

Category Scale, etc. Validity here refers to the closeness of the acuity level measured by nurses to patients' true acuity at the time of triage. In statistical language, we say that patient acuity level measured according to the systems is subject to error.

It was also pointed out that there is not a gold standard by which the true patient acuity could be measured. In practice, often an algorithm is designed to calculate some index as a proxy to assess validity based on surrogate outcome markers, such as hospitalization, total length of stay in the ED, cost of the ED visit, intensive care unit admission and ED resource usage, etc. (See papers by M. R. Baumann and T. D. Strout [4] and by S. Gouin, J. Gravel, D. K. Amre and S. Bergeron [15].) M. van Veen and H. A. Moll wrote a good review paper [35] providing an overview on the matter of reliability and validity of triage systems. I. van der Wulp, M. E. van Baar and A. J. P. Schrijvers published [33] a simulation study, an example for our understanding from the perspective of medical science, to assess the reliability and validity of the MTS.

We give a simple illustration of misclassification in emergency medical service. A patient who had headache might be assigned with priority 4 as if the headache was non life-threatening according to a triage system of four levels. But, if in fact the headache was due to a blockage in blood flow in the brain which may ultimately lead to a stroke by which the patient requires immediate attention. In this case, the patient really belong to the relatively priority 1 category and the aim is to detect the real underlying health concern giving present symptoms. We will dedicate this chapter to discuss on some important issues of patient priority misclassification.

Zee and Theil [34] were the ones first to consider misclassification problem in priority queues for the case of $R = 2$ only. Suppose in an ED there are only two categories of patients. True urgent patients arrive to the system as a Poisson process with mean rate λ_1 are mistakenly assigned to be in the non-urgent group at random at a mean rate δ_{12}. Non-urgent patents arrive to the system as a Poisson process with mean rate λ_2 and are mistakenly assigned to be in the urgent group at random at a mean rate δ_{21}. Under the condition of misclassification, Zee and Theil showed that the overall expected wait time can still be better than the mean wait time of a system where patients are treated simply by FCFS when the following criteria

$$\begin{aligned} \left(\frac{\delta_{12}}{\lambda_1} + \frac{\delta_{21}}{\lambda_2}\right) &< 1 \\ \frac{1}{\mu_1} &< \tfrac{1}{\mu_2} \end{aligned}$$

are met and relying on an implicit assumption that the treatment time of each patient does not depend on their assigned treatment priority at triage but on the true patient acuity which can be correctly assessed by a physician.

They proposed a method to allocate patients with uncertain treatment priority from either category to a mixture group with treatment priority after the most urgent but before the less urgent type. Suppose urgent patients are classified into the mixture group at a rate of ϵ_1 and non-urgent patients at a rate of ϵ_2. Under the assumption that $\epsilon_1 > \delta_{12}$ and $\epsilon_2 > \delta_{21}$, the overall mean wait time under such allocation when misclassification exists can be no worse compare to the FCFS system.

Zee and Theil's model assumptions are very practical but restrictive, and their results are very insightful for a two-priority single-channel system. The extension of introducing mixture groups is possible in priority queueing systems with $R > 2$, but currently there are no published papers studying such proposal. Again, our analysis deviates from the line of Cobham's [7] method but resource to the more general approach of using the FMCI technique when misclassification exists in patient triage prior to treatment, compare to the case if misclassification does not occur in triage from the beginning.

Suppose all patients are of R treatment priority classes corresponding to their true acuity levels. Ideally, we would like the priority queue to operate in the condition that all true acuity level i patients are correctly assigned treatment priority i at triage and patients are treated in the order of their priority between priority groups. If misclassification occurs at triage, suppose with probability p_{ij} an arrived true acuity level i patient is assigned treatment priority j by triage having mean treatment time $1/\mu_{ij}$ for all $i, j \in \{1, \ldots, R\}$, then with probability $p_{ii} = 1 - \sum\limits_{j=1, j \neq i}^{R} p_{ij}$ an acuity level i patient is assigned the correct treatment priority i having mean treatment time $1/\mu_{ii}$. If the mean treatment times are not affected by misclassification, then $1/\mu_{ij} = 1/\mu_{ii}$ for all j.

5.0.3 Defining and Estimating Mean Treatment Times Under Misclassification for Priority Queues

Two-Priority Queues

Let $N_1(t)$ and $N_2(t)$ denote the total number of arrivals of true acuity level 1 (more urgent) and level 2 (less urgent), respectively, during $[0, t)$ assuming $N_1(t)$ and $N_2(t)$ are independent Poisson processes with rates $\lambda_1 t$ and $\lambda_2 t$, respectively. Assuming from $N_1(t)$, an arrived acuity level 1 patient is misclassified to be of treatment priority 2 at random with probability p_{12}, and from $N_2(t)$, an arrived acuity level 2 patient is misclassified to be of priority 1 at random with probability p_{21}. Let N_{ij} be the number of true acuity level i patients being classified to be of treatment priority j, $i = 1, 2$ and $j = 1, 2$.

Theorem 5.1. *$N_{ij}(t)$ are independent Poisson processes with rates $p_{ij}\lambda_i t$ where $p_{11} = (1 - p_{12})$ and $p_{22} = (1 - p_{21})$.*

By definition, we have $N_1(t) = N_{11}(t) + N_{12}(t)$ and $N_2(t) = N_{21}(t) + N_{22}(t)$. We define $N_{1e}(t) = N_{11}(t) + N_{21}(t)$ and $N_{1e}(t)$ is a Poisson process with rate $\lambda_{1e}t = p_{11}\lambda_1 t + p_{21}\lambda_2 t$ denoting the number of treatment priority 1 patients arrived by triage under misclassification, and $N_{2e}(t) = N_{12}(t) + N_{22}(t)$ a Poisson process with rate $\lambda_{1e}t = p_{12}\lambda_1 t + p_{22}\lambda_2 t$ with treatment priority 2 by triage. The letter 'e' in the subscript is used to identify variables being subject to the condition of misclassification through out the chapter.

Theorem 5.2. *Given N_{1e}, N_{21} is BIN(N_{1e}, δ_{21}) distributed, $\delta_{21} = \dfrac{p_{21}\lambda_2}{(1-p_{12})\lambda_1 + p_{21}\lambda_2}$, and given N_{2e}, N_{12} is BIN(N_{2e}, δ_{12}) distributed, $\delta_{12} = \dfrac{p_{12}\lambda_1}{p_{12}\lambda_1 + (1-p_{12})\lambda_2}$.*

δ_{21} can be interpreted as the proportion of acuity level 2 patients in the priority 1 group under misclassification, and similarly δ_{12} the proportion of acuity level 1 patients in the priority 2 group under misclassification.

Then the following relations are true:

$$E[N_{21}|N_{1e}] = \delta_{21}N_{1e}$$

$$E[N_{11}|N_{1e}] = N_{1e} - \delta_{21}N_{1e} = (1-\delta_{21})N_{1e}$$

$$E[N_{12}|N_{2e}] = \delta_{12}N_{2e}$$

$$E[N_{22}|N_{2e}] = N_{2e} - \delta_{12}N_{2e} = (1-\delta_{12})N_{2e}$$

Typically in an emergency hospital, we assume that patients arrived to the emergency department first are tagged with their treatment priority according to their acuity at the triage station, and we assume that the treatment priority assignment may be subject to error. If the *TRUE* treatment priority can be measured with respect to patient acuity, then the misclassification error rate p_{ij} can be estimated.

Each patient being admitted to the ED spends S_{ijk} amount of time in treatment before departure from the system. S_{ijk} denotes the treatment time spent in ER of the kth patient whose true service priority score is i and is assigned with priority j. If $j = i$, then the classification is correct. If $j \neq i$, misclassification has occurred. For $k = 1, 2, \ldots, N_{ij}$, if we assume S_{ijk} are i.i.d. random variables having a distribution

$F_i(s)$ with mean $E[S_{ijk}] = 1/\mu_i$, then we are assuming that the mean treatment time of true priority i patient is not affected by misclassification.

We denote

$$S_{\cdot 1} = \sum_{k=1}^{N_{11}} S_{11k} + \sum_{k=1}^{N_{21}} S_{21k} = \sum_{k=1}^{N_{1e}-N_{21}} S_{11k} + \sum_{k=1}^{N_{21}} S_{21k} \tag{5.0.1}$$

the total service time rendered to patients being classified having treatment priority 1, and similarly,

$$S_{\cdot 2} = \sum_{k=1}^{N_{12}} S_{12k} + \sum_{k=1}^{N_{22}} S_{22k} = \sum_{k=1}^{N_{12}} S_{12k} + \sum_{k=1}^{N_{2e}-N_{12}} S_{22k} \tag{5.0.2}$$

for treatment priority 2 patients.

From (5.0.1) and (5.0.2), we denote

$$\begin{aligned} \bar{S}_{\cdot 1} = & \ \tfrac{1}{N_{1e}} S_{\cdot 1} = & \tfrac{1}{N_{1e}} \left(\sum_{k=1}^{N_{1e}-N_{21}} S_{11k} + \sum_{k=1}^{N_{21}} S_{21k} \right) \\ \bar{S}_{\cdot 2} = & \ \tfrac{1}{N_{2e}} S_{\cdot 2} = & \tfrac{1}{N_{2e}} \left(\sum_{k=1}^{N_{12}} S_{12k} + \sum_{k=1}^{N_{2e}-N_{12}} S_{22k} \right) \end{aligned} \tag{5.0.3}$$

the average patient time spent in ER by the misclassification.

In application, N_{1e} and N_{2e} are observable. If we reasonably assume that N_{1e}, N_{2e}, N_{12} and N_{21} are independent of S_{ijk}, then by the law of total probability, expectation of the relations in (5.0.3) are

$$\begin{aligned} \frac{1}{\mu_{1e}} \overset{\text{def}}{=} E\left[\bar{S}_{\cdot 1}|N_{1e}\right] &= E\left[\tfrac{1}{N_{1e}}\left(\sum_{k=1}^{N_{1e}-N_{21}} S_{11k} + \sum_{k=1}^{N_{21}} S_{21k}\right)\middle|N_{1e}\right] \\ &= E_{N_{21}|N_{1e}}\left\{E\left[\tfrac{1}{N_{1e}}\left(\sum_{k=1}^{N_{1e}-N_{21}} S_{11k} + \sum_{k=1}^{N_{21}} S_{21k}\right)\middle|N_{21}\right]\right\} \\ &= E_{N_{21}|N_{1e}}\left\{\tfrac{(N_{1e}-N_{21})}{N_{1e}}E[S_{11k}] + \tfrac{N_{21}}{N_{1e}}E[S_{21k}]\right\} \qquad (5.0.4) \\ &= \frac{(1-\delta_{21})N_{1e}}{N_{1e}}\frac{1}{\mu_1} + \frac{\delta_{21}N_{1e}}{N_{1e}}\frac{1}{\mu_2} \\ &= (1-\delta_{21})\frac{1}{\mu_1} + \delta_{21}\frac{1}{\mu_2} \end{aligned}$$

and similarly

$$\frac{1}{\mu_{2e}} \overset{\text{def}}{=} E\left[\bar{S}_{\cdot 1}|N_{2e}\right] = \delta_{12}\frac{1}{\mu_1} + (1-\delta_{12})\frac{1}{\mu_2} \qquad (5.0.5)$$

Note, the above two equations do not work well if δ_{12} and δ_{21} are 0.5, that which the two equations are linearly dependent and there are infinitely many solutions of $\frac{1}{\mu_1}$ and $\frac{1}{\mu_2}$.

If δ_{21} and δ_{12} are not known and $1/\mu_1$ and $1/\mu_2$ are to be estimated, the solution of equations (5.0.4) and (5.0.4) depends on the estimation of δ_{21}, δ_{12}, $1/\mu_{1e}$ and $1/\mu_{2e}$. Naturally we propose to use the estimates $\hat{\delta}_{21} = n_{21}/n_{1e}$ and $\hat{\delta}_{12} = n_{12}/n_{2e}$ with

$$\widehat{\frac{1}{\mu_{1e}}} = \frac{1}{n_{1e}}\left(\sum_{k=1}^{n_{11}} s_{11k} + \sum_{k=1}^{n_{21}} s_{21k}\right)$$

and

$$\widehat{\frac{1}{\mu_{2e}}} = \frac{1}{n_{2e}}\left(\sum_{k=1}^{n_{12}} s_{12k} + \sum_{k=1}^{n_{22}} s_{22k}\right).$$

On the other hand, in a given sample $1/\mu_1$ can also be estimated by calculating the sample average treatment time of patients re-assessed of true acuity score 1 who are under correct triage and those of re-assessed true acuity score 1 assigned to treatment priority 2 group,

$$\widehat{\frac{1}{\mu_1}} = \frac{1}{n_{11}+n_{12}}\left(\sum_{k=1}^{n_{11}} s_{11k} + \sum_{k=1}^{n_{12}} s_{12k}\right)$$

Similarly, $1/\mu_2$ can be estimated by

$$\widehat{\frac{1}{\mu_2}} = \frac{1}{n_{21}+n_{22}}\left(\sum_{k=1}^{n_{21}} s_{21k} + \sum_{k=1}^{n_{22}} s_{22k}\right).$$

Then $1/\mu_{1e}$ and $1/\mu_{2e}$ can be estimated using equations (5.0.4) and (5.0.5) by replacing δ_{ij} with $\hat{\delta}_{ij}$ and $1/\mu_i$ with $\widehat{1/\mu_i}$.

If $1/\mu_{12}$ differs from $1/\mu_1$ and $1/\mu_{21}$ differs from $1/\mu_2$ under the influence of misclassification, then equations (5.0.4) and (5.0.5) should be restated as

$$\begin{aligned} \frac{1}{\mu_{1e}} &\stackrel{\text{def}}{=} (1-\delta_{21})\frac{1}{\mu_{11}} + \delta_{21}\frac{1}{\mu_{21}} \\ \frac{1}{\mu_{2e}} &\stackrel{\text{def}}{=} \delta_{12}\frac{1}{\mu_{12}} + (1-\delta_{12})\frac{1}{\mu_{22}} \end{aligned} \qquad (5.0.6)$$

For each $i = 1, 2$ and $j = 1, 2$, $1/\mu_{ij}$ can be estimated by

$$\widehat{\frac{1}{\mu_{ij}}} = \frac{1}{n_{ij}}\sum_{k=1}^{n_{ij}} s_{ijk}$$

and then $1/\mu_{1e}$ and $1/\mu_{2e}$ of equations in (5.0.6) should be estimated by replacing $1/\mu_{ij}$ with $\widehat{1/\mu_{ij}}$.

Extension to More-than-Two Priority Model Under Misclassification

Suppose now arrivals are of r categories. In a fixed time interval $[0, t)$, for some large enough t, category i, $i = 1, \ldots, R$, patients arrive to the ED according to a

Poisson process at a mean rate of $\lambda_i t$. Let $N(t)$ denote the total number of arrivals during $[0, t)$, δ_{ij} be the rate at which true acuity level i patients are classified to be of priority j, and $N_{ij}(t)$ be the number of true service priority i patients being classified to be of treatment priority j, $i = 1, \ldots, R$ and $j = 1, \ldots, R$. Let $N_{je} = N_{1j}+N_{2j}+\cdots+N_{Rj}$ denote the number of treatment priority j patientss arrived under the condition of misclassification and is observable for all j in practice. Theorem 5.1 applies here.

Theorem 5.3. *Condition on that $N_{je}, N_{1j}, \ldots, N_{Rj}$ are independent binomial random variables, $N_{ij} \sim BIN(N_{je}, \delta_{ij})$, $i = 1, \ldots, R$, for $j = 1, \ldots, R$ where*

$$\delta_{ij} = \begin{cases} \frac{p_{ij}\lambda_i}{\left(1-\sum\limits_{k=1,k\neq j}^{R} p_{jk}\right)\lambda_j+\sum\limits_{k=1,k\neq j}^{R} p_{kj}\lambda_k} & \text{for } i \neq j, \\ 1 - \sum\limits_{i=1,i\neq j}^{R} \delta_{ij} & \text{for } i = j. \end{cases}$$

It follows that

$$E\left[N_{ij} \mid N_{je}\right] = \begin{cases} \delta_{ij} N_{je} & \text{for } i \neq j, \\ \left(1 - \sum\limits_{i=1,i\neq j}^{R} \delta_{ij}\right) N_{je} & \text{for } i = j. \end{cases}$$

In hospital service, if ER patients were triaged upon arrival and the triage score were re-determined post treatment to serve as the true patient acuity scale, the score re-determined post treatment may be concordant or discordant to the triage score measured upon arrival. δ_{ij} has the interpretation as the expected proportion of patients with post treatment score i in an observed sample of patients with triage score

j measured at arrival.

For models of more than two service priorities, equations in (5.0.6) can easily be extended as

$$\frac{1}{\mu_{je}} \stackrel{\text{def}}{=} E\left[\bar{S}_{\cdot 1} | N_{je} = n_{je}\right] = \left(1 - \sum_{i=1, i\neq j}^{R} \delta_{ij}\right)\frac{1}{\mu_{jj}} + \sum_{i=1, i\neq j}^{R} \delta_{ij}\frac{1}{\mu_{ij}} \tag{5.0.7}$$

for $j = 1, \ldots, R$ in general. For each $i = 1, \ldots, R$ and $j = 1, \ldots, R$, $1/\mu_{ij}$ can be estimated by

$$\widehat{\frac{1}{\mu_{ij}}} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} s_{ijk} \tag{5.0.8}$$

and δ_{ij} by

$$\hat{\delta}_{ij} = \frac{n_{ij}}{n_{je}} \tag{5.0.9}$$

for all i and j.

Then, $1/\mu_{je}$ in equation (5.0.7) can be estimated by replacing δ_{ij} with $\hat{\delta}_{ij}$ and $1/\mu_{ij}$ with $\widehat{1/\mu_{ij}}$. For the special case under the assumption that mean treatment times are not affected by misclassification, we have $1/\mu_{ij} = 1/\mu_i$ for all j that can be estimated by

$$\widehat{1/\mu_i} = \frac{1}{(n_{i1} + \cdots + n_{iR})} \sum_{j=1}^{R} \sum_{k=1}^{n_{ij}} s_{ijk}$$

in an observed sample data.

For example, in a larger hospital centre, there may be up to 200 patients visiting ER in a typical week. Suppose patients were triaged on an acuity scale of three to determine their treatment priority. We have $n_{1e} + n_{2e} + n_{3e} = 200$. From n_{je} for each $j = 1, 2, 3$, if time and resources permits, re-assess each patient's true acuity level post treatment to determine n_{ij} for all j. Estimates $\widehat{1/\mu_{ij}}$ and $\hat{\delta}_{ij}$ can be obtained as in equations (5.0.8) and (5.0.9) for $i, j = 1, 2, 3$.

When a large data set is available, one approach to estimate the δ_{ij}'s is to use the stratified sampling technique to estimate the misclassification rates in each treatment priority category. At the same time, estimates of $1/\mu_{ij}$'s can be obtained.

With respect to the nurse triage score, the weighted mean treatment time of patients with triage score j in an observed sample is

$$\frac{1}{\mu_{je}} = \left(1 - \sum_{i=1, i \neq j}^{R} \delta_{ij}\right) \frac{1}{\mu_{jj}} + \sum_{i=1, i \neq j}^{R} \delta_{ij} \frac{1}{\mu_{ij}}, \tag{5.0.10}$$

or is

$$\frac{1}{\mu_{je}} = \left(1 - \sum_{i=1, i \neq j}^{R} \delta_{ij}\right) \frac{1}{\mu_{j}} + \sum_{i=1, i \neq j}^{R} \delta_{ij} \frac{1}{\mu_{i}}, \tag{5.0.11}$$

depending on whether or not misclassification affects the mean treatment times of true acuity i patients.

To study the effects of misclassification on waiting times in our priority queueing models as in Chapters 3 and 4, we simply replace known parameters of $1/\mu_j$ by $\widehat{1/\mu_{je}}$ calculated using (5.0.10) or (5.0.11) in the imbedding procedure when constructing tpm's. Comparing to the system behaviour and the waiting-time distribution under the condition that the triage process was perfectly conducted, then we obtain the waiting time distributions replacing the parameters $1/\mu_j$ by $\widehat{1/\mu_{jj}}$ from (5.0.10), or simply by $\widehat{1/\mu_j}$ from (5.0.11), in the imbedding procedure when constructing tpm's. When comparing models with and without priority misclassification, numerical studies can be used to devise some kind of tolerance on rates of misclassification so that mean wait times do not get prolonged over a threshold from the effect of error in priority assignment.

Chapter 6

Conclusion and Some Discussions

With the help of the finite Markov chain imbedding technique, we are now able to model $M/M/1$ preemptive repeat-different and non-preemptive priority queues with thresholds on the maximum number of customers allowed to be in service and in the queueing system. In this thesis, we consider only queueing processes having two major assumption that customers arrive to the system as Poisson processes and the service-time distribution being exponential. The same technique can definitely be used to model priority queues beyond the scope of the aforementioned two restrictive assumptions. One can see that the waiting time distribution of a queueing process can easily be obtained as the waiting time distribution of a set of particular events or patterns of an imbedded Markov chain which portrays the original queueing process. Since a Markov process is completely characterized by its transition probability matrix and an initial state, a priority queueing system can be characterized by a Markov process with sufficient information of the system known. The imbedding procedure induces a finite state space consists of all possible events or realizations of a Markov process over a discretised time horizon. Properly imposed state transition rules defined by the queueing discipline together with determined state transition probabilities and an initial state distribution, can be custom selected, is required to calculate the waiting time distribution. Next, we provided a few future research topics which can be considered in the last section of this thesis.

6.1 Extensions to other Queues and Research Topics Envisaged

Priority queueing models with the following features have real applications in queueing theory, various priority scheduling to control/reduce wait-time of a queue, emergency department service system, reliability theory, repairing systems and police and fire fighting rescue aiming to reduce some measure of casualty, etc.:

- allowing multiple departures and therefore allowing also from the buffer zone to have multiple waiting customers entering service within an interval $[t - \Delta t, t)$, for any t;

- consider $M/M/1$-$r/b/c_1, \ldots, c_R$ model with service thresholds c_i on class i customers for $i = 1, \ldots, R$;

- possible priority jump, particularly a customer from class j to class i, $j > i$, allowing reassessment of patient status over some interval of time and therefore making patient priority reassignment possible;

- other priority queues having other than Poisson arrival process or exponential service assumptions;

- multi-channel priority queues;

- various priority scheduling such as preemptive-resume and preemptive repeat-identical, etc.

Bibliography

[1] M. A. Aczel. The effect of introducing priorities. *Journal of the Operations Research Society of America*, 8(5):730–733, 1960.

[2] Attahiru S. Alfa. Matrix-geometric solution of discrete time map/ph/1 priority queue. *Naval research logistics*, 45(1):23–50, February 1998.

[3] G. Anderson, C. Black, E. Dunn, J. Alonso, J. Christian-Norregard, T. Folmer-Anderson, and P. Bernth-Peterson. Willingness to pay to shorten waiting time for cataract surgery. *Health Affairs*, 16(5):181–190, 1997.

[4] M. R. Baumann and T. D. Strout. Evaluation of the emergency severity index (version 3) triage algorithm in pediatric patients. *Academic Emergency Medicine*, 12(3):219–224, March 2005.

[5] A. Bedford and P. Zeephongsekul. On a dual queueing system with preemptive priority service discipline. *European Journal of Operational Research*, 161(1):224–239, February 2005.

[6] CIHI. Understanding emergency wait times: How long do people spend in emergency departments in ontario? Report, Canadian Institute for Health Information, January 2007.

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

[8] A. Cobham. Priority assignment - a correction. *Journal of the Operations Research Society of America*, 3(4):547, 1955.

[9] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1 of *Wiley Series in Probability and Mathematical Statistics*. John Wiley and Sons, Inc., 3rd edition, 1968.

[10] J. C. Fu and M. V. Koutras. Distribution theory of runs: a markov chain approach. *J. Am. Stat. Assoc.*, 89:1050–1058, 1994.

[11] James C. Fu. Exact and limiting distributions of the number of successions in a random permutation. *Annals of the Institute of Statistical Mathematics*, 47(3):435–446, September 1995.

[12] James C. Fu. Distribution of the scan statistic for a sequence of bistate trials. *Journal of Applied Probability*, 38(4):908–916, December 2001.

[13] James C. Fu and W.Y. Wendy Lou. *Distribution Theory of runs and Patterns and*

Its Application: A Finite Markov Chain Imbedding Approach. World Scientific Publishing Co. Pte. Ltd., 2003.

[14] James C. Fu, Fred A. Spiring, and Hansheng Xie. On the average run lengths of quality control schemes using a markov chain approach. *Statistics and Probability Letters*, 56:369–380, 2002.

[15] S. Gouin, J. Gravel, D. K. Amre, and S. Bergeron. Evaluation of the paediatric canadian triage and acuity scale in a pediatric ed. *The American Journal of Emergency Medicine*, 23(3):243–247, May 2005.

[16] J. L. Holley. Waiting line subject to priorities. *Journal of the Operations Research Society of America*, 2(3):341–343, 1954.

[17] D. Isaacson and R. Madsen. Positive columns for stochastic matrices. *Journal of Applied Probability*, 11(4):829–835, Dec. 1974.

[18] N. K. Jaiswal. *Priority Queues*, volume 50 of *Mathematics in science and engineering*. New York : Academic Press, 1968.

[19] David G. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society*, 13(2):151–185, January 1951.

[20] David G. Kendall. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354, Sept. 1953.

[21] H. Kesten and J. Th. Runnenberg. Priority in waiting line problems. *Proc Koninklijke Nederlandse Akademie van Wetnchappen*, A60(3):312–336, 1957.

[22] A. Y. Khintchine. Mathematisches uber die erwartung vor einem offentlichen schal-ter. *Mat. Sbornik*, 39(4):73–84, 1932.

[23] Peter Lancaster and Miron Tismenetsky. *The Theory of Matrices.* Computer Science and Applied Mathematrics. Academic Press, Inc., second edition with applications edition, 1985.

[24] Hui Li and Yiqiang Q. Zhao. Exact tail asymptotics in a priority queue - characterizations of the preemptive model. *Queueing Syst.*, 63:355–381, 2009.

[25] Grégory Nuel. Effective p-value computations using finite markov chain imbedding (fmci): application to local score and to pattern statistics. *Algorithms for Molecular Biology*, 1(1):5, April 2006.

[26] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*, november 14, 2008 edition, November 2008.

[27] F. Pollaczek. Uber eine aufgabe dev wahrscheinlichkeitstheorie. *Mathematische Zeitschrift*, 32:64–100 and 729–850, 1930.

[28] Jr. R. G. Miller. Priority queues. *Annals of Mathematical Statistics*, 31(1):86–103, 1960.

[29] Karuna Ramachandran. *Matrix Geometric Methods in Priority Queues.* Doctor of philosophy thesis, The University of Western Ontario, July 1997.

[30] D. W. Spaite, F. Bartholomeaux, J. Guistoand E. Lindberg, B. Hull, A. Eyherabide, S. Lanyon, E. A. Criss, T. D. Valenzuela, and C. Conroy. Rapid process redesign in a university-based emergency department: Decreasing waiting time intervals and improving patient satisfaction. *Annals of Emergency Medicine*, 39(2):168–177, 2002.

[31] David A. Stanford. Waiting and interdeparture times in priority queues with poisson- and general-arrival streams. *Operations Research*, 45(5):725–735, Sept.-Oct. 1997.

[32] H. M. Taylor and S. Karlin. *An Introduction To Stochastic Modeling.* Academic Press, 3rd edition, 1998.

[33] I. van der Wulp, M. E. van Baar, and A. J. P. Schrijvers. Reliability and validity of the manchester triage system in a general emergency department patient population in the netherlands: results of a simulation study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 25:431–434, 2008.

[34] S. P. van der Zee and H. Theil. Priority assignment in waiting-line problems under conditions of misclassification. *Operations Research*, 9(6):875–885, Nov.-Dec. 1961.

[35] Mirjam van Veen and Henriette A. Moll. Reliability and validity of triage systems in paediatric emergency care. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 17:38, August 2009.

[36] Dietmar Wagner. Analysis of a finite capacity multi-server model with non-preemptive priorities and non-renewal input. In S. R. Chakravarthy and A. S. Alfa, editors, *Matrix-Analytic Methods in Stochastic Models*, Lecture Notes in Pure and Applied Mathematics, pages 67–86. CRC Press, September 1996.

[37] Dietmar Wagner. Waiting times of a finite-capacity multi-server model with non-preemptive priorities. *European Journal of Operational Research*, 102(1):227–241, October 1997.

[38] Dietmar Wagner and Udo R. Krieger. Analysis of a finite buffer with non-preemptive priority scheduling. *Communications in Statistics - Stochastic Models*, 15:345–365, 1999.

[39] Jingui Xie, Qi-Ming He, and Xiaobo Zhao. On the stationary distribution of queue lengths in a multi-class priority queueing system with customer transfers. *Queueing Syst. Theory Appl.*, 62:255–277, July 2009.

[40] P. Zeephongsekul and A. Bedford. Waiting time analysis of the multiple priority dual queue with a preemptive priority service discipline. *European Journal of Operational Research*, pages 886–908, 2006.

[41] Hongbo Zhang and Dinghua Shi. Explicit solution for m/m/1 preemptive priority queue. *International Journal of Information and Management Sciences*, 21:197–208, 2010.